1-1-2015

Single Human Cells Use Transcriptional Mechanisms to Compensate for Differences in Cell Size and DNA Content

Olivia Padovan-Merhar
University of Pennsylvania, olivia.padovan@gmail.com

Single Human Cells Use Transcriptional Mechanisms to Compensate for Differences in Cell Size and DNA Content

Abstract
Human cells are dynamic: they grow, replicate their genetic information (DNA), and divide. Clonal populations of cells can display marked heterogeneity in size, leading to significant variability in the ratio of DNA to cellular volume. Despite this variability, cells must maintain a constant concentration of RNA and protein, produced from DNA, to ensure proper functionality. How do larger cells produce more output from the same amount of DNA? How do cells that have replicated their DNA prior to cellular division produce the same output as before? Using RNA fluorescence in situ hybridization (RNA FISH), we visualize and count individual RNA molecules in single cells, allowing for precise quantification of transcriptional output of single genes. We also use single-cell RNA sequencing to quantify transcriptional output from all ~20,000 genes encoded in the genome simultaneously. Surprisingly, we discovered that the cell implements two separate transcriptional mechanisms to compensate for changes in cell size and DNA content. Through cell-fusion experiments, we show that a diffusible trans factor, which we believe may be RNA polymerase II, increases transcriptional burst size in larger cells, compensating for changes in volume. Meanwhile, a DNA-linked cis-acting factor reduces the frequency of transcription per gene copy by a factor of two upon DNA replication, allowing the cell to still produce the same amount of RNA after replication, despite having twice the number of DNA copies. We show that transcription depends strongly on volume, and we therefore present a new "noise measure" which provides a measure of gene expression variability that takes volume into account. We perform single-cell RNA sequencing to measure noise genome-wide, and find that cell-type-specific genes tend to exhibit more expression noise than genes that are ubiquitously expressed across cell types. Finally, we have uncovered a fundamental mechanism by which cells are able to functionally compensate for naturally-occurring variability in size and DNA copy number.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Physics & Astronomy

First Advisor
Arjun Raj

Keywords
Cell volume, Concentration, DNA, RNA, Transcription

Subject Categories
Cell Biology | Molecular Biology | Physics

SINGLE HUMAN CELLS USE TRANSCRIPTIONAL MECHANISMS TO COMPENSATE FOR DIFFERENCES IN CELL SIZE AND DNA CONTENT

Olivia M. Padovan-Merhar

A DISSERTATION

in

Physics and Astronomy

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2015

Supervisor of Dissertation

_____

Arjun Raj

Assistant Professor of Bioengineering

Graduate Group Chairperson

_____

Marija Drndić, Professor of Physics and Astronomy

Dissertation Committee

Mark Goulian, Professor of Physics and Astronomy and Professor of Biology

Philip C. Nelson, Professor of Physics and Astronomy

Alison M. Sweeney, Assistant Professor of Physics and Astronomy

Vijay Balasubramanian, Professor of Physics and Astronomy

SINGLE HUMAN CELLS USE TRANSCRIPTIONAL MECHANISMS TO COMPENSATE FOR DIFFERENCES IN CELL SIZE AND DNA CONTENT

Acknowledgements

To Raj Lab members:

I'll be hard pressed to find a

greater group than you

Arjun, great PI:

You are the best mentor I

could have imagined

Arjun gets two lines.

You helped turn me into a

real biologist

To Gautham Nair:

You have a critical eye

Thanks for thoughtful chats

To undergrads and

rotation students: Thanks for

all of your hard work!

Jeff Carey the great:

You came up with an awesome

experiment. Yay!

Stirling and Abhi:

You helped us do things we could

not do on our own

Committee members:

Thank you for your helpful thoughts

and thesis guidance

WPS:

All of you are amazing

Dancing with you rocks

Raja YTTs:

You keep me sane and centered

Keep belly breathing

To my family:

Thank you for love and support

I love all of you

To Derek, partner

in life and dance. Can't wait for

our next adventure!

ABSTRACT

SINGLE HUMAN CELLS USE TRANSCRIPTIONAL MECHANISMS TO COMPENSATE FOR DIFFERENCES IN CELL SIZE AND DNA CONTENT

Olivia M. Padovan-Merhar

Arjun Raj

Human cells are dynamic: they grow, replicate their genetic information (DNA), and divide. Clonal populations of cells can display marked heterogeneity in size, leading to significant variability in the ratio of DNA to cellular volume. Despite this variability, cells must maintain a constant concentration of RNA and protein, produced from DNA, to ensure proper functionality. How do larger cells produce more output from the same amount of DNA? How do cells that have replicated their DNA prior to cellular division produce the same output as before? Using RNA fluorescence *in situ* hybridization (RNA FISH), we visualize and count individual RNA molecules in single cells, allowing for precise quantification of transcriptional output of single genes. We also use single-cell RNA sequencing to quantify transcriptional output from all (∼20,000) genes encoded in the genome simultaneously. Surprisingly, we discovered that the cell implements two separate transcriptional mechanisms to compensate for changes in cell size and DNA content. Through cell-fusion experiments, we show that a diffusible *trans* factor, which we believe may be RNA polymerase II, increases transcriptional burst size in larger cells, compensating for changes in volume. Meanwhile, a DNA-linked *cis*-acting factor reduces the frequency of transcription per gene copy by a factor of two upon DNA replication, allowing the cell to still produce the same amount of RNA after replication, despite having twice the number of DNA copies. We show that transcription depends strongly on volume, and we therefore present a new "noise measure" which provides a measure of gene expression variability that takes volume into account. We perform

single-cell RNA sequencing to measure noise genome-wide, and find that cell-type-specific genes tend to exhibit more expression noise than genes that are ubiquitously expressed across cell types. Finally, we have uncovered a fundamental mechanism by which cells are able to functionally compensate for naturally-occurring variability in size and DNA copy number.

Contents

List of Figures

Chapter 1

Introduction

The human body harbors over 200 different types of cells, each of which has an individual, specialized function. Accordingly, there is a range of human cell sizes that spans many orders of magnitude. Interestingly, even cells of the same type can show large size variability both *in vivo* and when grown in culture [7, 11, 67]. Despite this large variability in volume, we assume that clonal populations of cells in culture maintain similar functionality, suggesting that something must be constant between these cells, even if volume is not. The rates of biochemical reactions within a cell depend on the concentrations of reactants and enzymes, suggesting that in order for large and small cells to function similarly, the concentrations of biomolecules within the cells must remain constant, and hence the absolute numbers of molecules would have to scale roughly linearly with cellular volume (see Marguerat et al. for an excellent review [34]). How does the cell tackle this problem?

1.1 Gene expression and cellular volume

Yeast, bacteria, and many plants display a striking correlation between cell (or organism) size and the number of copies of DNA within each cell [38, 40, 52, 66]. Yeast double in size after DNA replication as they progress from the G1 to G2 phases of the cell cycle [80], and haploid species of yeast are, on average, half the size of their diploid

counterparts [75]. Many agricultural crops that are grown for their size and grain output have higher ploidy than the same plants grown in the wild [10, 44, 52, 72]. Yet mammalian cells do not always follow this trend. Surprisingly, DNA is the one molecule in mammalian cells that need not scale with cell size, either within a population of clonal cells, or between cell types of different sizes and functions. While human cells are, on average, larger in G2 than G1, volumes in all stages of the cell cycle are highly variable. In most mammalian cells, DNA will typically exist in two or four copies per cell, and even cells with the same number of DNA molecules can differ widely in size [7, 67, 78]; as such, DNA concentration can vary dramatically from cell to cell. This poses a challenge for cellular homeostasis, for if two otherwise identical cells with the same DNA content had different volumes, then the larger cell must somehow maintain a higher number of biomolecules with that same amount of DNA.

Biologists have been curious about this puzzle for decades, beginning in the 1970s with papers from Crissman and Steinkamp [11] and Fraser and Nurse [19]. Crissman and Steinkamp performed a primarily observational study, where they used flow cytometry to measure cellular volume and total protein content simultaneously in populations of human cells. They found that the distribution of ratios of protein-to-volume was much tighter than either the DNA-to-volume or DNA-to-protein ratios, giving us our first hint that (1) cells display a variety of volumes even when grown in culture, and (2) cells employ some mechanism to maintain protein at a constant concentration, regardless of cell size. In 1979, Fraser and Nurse began to unravel the transcriptional underpinnings of the scaling of cellular components with volume, using yeast as a model organism. They measured RNA concentration (RNA/volume) for three strains that had different "gene concentrations" (DNA/volume ratios), and found that, despite a 2-fold change in gene concentration between the largest and smallest strain, the change in mRNA concentration was negligible. They concluded that the cell must have a compensatory

transcriptional mechanism in place to maintain a constant concentration of RNA despite changes in gene concentration, and provided evidence that smaller cells produce less RNA by delaying an increase in transcription until late in G2. Even with the relatively limited techniques available nearly 40 years ago, biologists were able to detect this scaling mismatch between DNA and volume and begin to understand how cells are able to compensate for it through transcription. However, these measurements were primarily observational and were either performed on bulk populations of cells or were relatively qualitative. These studies began to define the problem—how can some cells produce more RNA from the same amount of DNA?—but they were not able to definitively answer that question.

It was not until relatively recently that the advent of RNA sequencing, single-molecule imaging, and other technological developments made it possible to gain new understanding of transcriptional regulation and its relationship to cell size. A number of recent studies have shown that both the amount of RNA and protein generally scales with cellular volume [34, 35, 55, 71, 79] and ploidy [75], with some further finding that transcription itself changes in mutants with larger or smaller cell volumes [19, 55, 79]. Most of these studies were performed in yeast, with a few notable exceptions [39, 55, 71].

Schmidt and Schibler [55] played a major role in establishing this field by defining "cell size regulation", or the mechanism by which cells of different sizes are able to produce different amounts of RNA from the same amount of DNA. They observed that cells in different human tissues had different DNA/cytoplasm ratios and that total RNA scaled with cytoplasmic size, not DNA content. For the first time in human cells, they showed that transcription rate, not degradation rate, scaled with cell size. However, they studied transcriptional differences between cell types, and were therefore unable to make a claim about how cells from a single lineage compensate for natural

variability in volume.

Miettinen and colleagues [39] used a clever method to examine cells of different sizes within the same tissue, by studying regenerated mouse liver tissue with and without a particular gene knockout. Wild-type livers regenerated by producing many new cells, while knockout livers regenerated by increasing the size of existing cells, providing the researchers with large and small liver cells to examine. They performed RNA sequencing on these large and small cells to identify differentially expressed genes, but found that most genes were not differentially expressed when they normalized to the total number of reads. We note here that by normalizing to the total number of reads, one is essentially normalizing away any volume effects and will not observe any global changes in expression between conditions. Therefore it is likely that the majority of genes in the genome were in fact expressed at higher global levels in the larger cells, but all genes were upregulated by the same amount. This finding lends further credence to our emerging hypothesis that there exists a global control that can change expression levels of all genes in a concordant manner to scale with cell size.

In another interesting work, Watanabe and colleagues [71] discovered a small *C. elegans* mutant that contained the same number of cells as the larger wild-type worm, and showed that the total protein content of worms from both strains was approximately proportional to organism size. All three of these studies further show that there is a scaling of RNA and protein with volume within multicellular eukaryotic organisms. However, these studies also demonstrate the difficulty of making significant progress on this mechanistic question in higher eukaryotes, due both to the difficulty of performing perturbative experiments in such cells, and the challenges associated with studying single cells as opposed to bulk populations.

On the other hand, there are a few very interesting mechanistic papers on this subject in yeast. Wu and colleagues [75] compared large tetraploid yeast cells to mutant

haploid cells of the same size, and found that their gene expression profiles were similar. Furthermore, they found that gene expression increased with cell size by examining haploid mutants of different sizes. These results imply that cells have a mechanism allowing gene expression to scale with volume, independent of DNA content. Zhurinsky and colleagues [79] took this a step further, generating cells that spanned a five-fold range of DNA/protein ratios. Within a certain "biological" window, they found the same results as Wu et al., namely that gene expression scaled with volume. However, they pushed the limits of the DNA/protein ratio, finding that in very large cells, DNA content became limiting and they observed a plateau in expression levels. Moreover, they examined levels of individual transcripts, and found that for all cells, nearly all genes were expressed at similar relative levels in all conditions, again demonstrating the existence of a global transcriptional regulator. This is the first piece of literature to concretely begin to define a "limiting factor" linking cell volume to transcription.

All of the above experiments have been extremely influential in defining this field and addressing key problems, however, there are still many unanswered questions. For example, studying different tissue types or using mutants does not allow one to establish a causal relationship between cellular volume changes and transcript abundance. Such a relationship would have strong implications for the interpretation of gene expression measurements because if cellular volume changes can in and of themselves change global expression levels, observations of changes in global expression levels in response to various perturbations may actually be the indirect consequence of changes to cellular volume rather than resulting from direct global transcriptional responses to the perturbations *per se*. Further, while these experiments have hinted at the existence of a limiting factor linking transcription and volume, it has not been fully characterized. Nor do we understand exactly how transcription is modulated in cells of different sizes, or how cells compensate for changes in DNA in addition to

volume.

1.2 Stochastic gene expression

All of the literature we have discussed so far points to the existence of a mechanism allowing cells to compensate for changes in volume and DNA content in order to maintain a constant concentration of RNA and protein, ostensibly within both bulk populations and single cells. Yet there exists an enormous body of work suggesting that transcription is inherently stochastic, and that RNA and protein levels vary widely at the single-cell level [50, 51, 54]. We will discuss this literature and see if it is possible to reconcile these seemingly contradictory findings.

On the face of it, transcription must be random to some extent. To simply initiate transcription, a whole host of transcription factors and cofactors must, through stochastic diffusion, arrive at an accessible promoter and assemble into a transcription initiation complex. In principle, there are two ways in which transcription can proceed. One possibility is that "active" genes are constitutively "ON", and the rate-limiting step of transcription is the time it takes for all of the correct elements to arrive at the promoter. This would cause RNA to be produced in a stochastic manner, but with a uniform probability per unit time to produce a single molecule of RNA. Such production would lead to a Poisson distribution of RNA in single cells. However, significant evidence in bacteria, yeast, and higher eukaryotes suggests that RNA production is better approximated by a two-state model [28], in which the promoter switches stochastically between "ON" and "OFF" states. In this model, multiple RNA molecules are produced in short bursts during the "ON" state, leading to a much broader distribution in RNA counts than in the Poisson case. This is referred to as "bursty transcription".

A few key papers demonstrated the existence of bursty transcription in bacteria and eukaryotes. Some studies used a live reporter assay to observe transcription in real time, while others were able to infer dynamics through analysis of static RNA distributions in fixed cells. Live-cell RNA imaging is primarily done through the use of the MS2 reporter assay, in which two constructs are introduced into a cell: a gene of interest attached to a cassette of MS2 binding sites, and a gene encoding the MS2 protein fused to a fluorescent protein. By counting instances (or summing total fluorescence intensity) of the reporter fluorescence, one can read out the total amount of the RNA of interest in the cell. Golding and colleagues [20] used this method to observe transcription of a fluorescent protein transgene in *E. coli*. They observed that transcription occurred in bursts, and that the gene appeared to fluctuate between an "ON" and an "OFF" state, providing some of the first live-cell evidence for transcriptional bursting. Chubb and colleagues [8] used a similar methodology to observe transcription of an endogenous gene in *Dictyostelium*. By using targeted integration, they inserted a cassette of MS2 binding sites directly into a gene, and used a fluorescently-tagged MS2 protein to read out transcriptional activity. They found that transcription of this endogenous gene was also pulsatile, showing for the first time that endogenous loci—not just transgenes—exhibit bursty transcription in eukaryotes.

More recently, Suter and colleagues [59] engineered a system to study transcriptional dynamics in live mammalian cells. They inserted the coding sequence for a short-lived luciferase protein (encoded by a short-lived RNA) downstream of either endogenous or synthetic promoters. They argued that because the reporter and its RNA are both short-lived, readout from the reporter should be indicative of transcriptional dynamics. For all types of promoters used, the authors found that transcription was bursty. These reports provide undeniable evidence that transcription occurs in

bursts, be it in bacteria, single-celled eukaryotes, or mammals. However, most of these studies suffer from a significant drawback, namely that most of these observations are performed for transgenes, which are likely not subject to the same sorts of regulation as endogenous genes. The one study that did observe transcription of an endogenous gene [8] did not use an easily-scalable technique, therefore making it difficult to study multiple genes and draw any broad conclusions about transcription in general. To fully understand transcriptional regulation as it occurs naturally, we need to be able to study transcriptional output from an unperturbed gene locus. We are interested in discovering whether the transcriptional patterns of endogenous genes are coordinated in such a way to allow RNA and protein to scale with cellular volume.

RNA fluorescence *in situ* hybridization (RNA FISH) is a technique that allows one to visualize individual RNA molecules in single cells using a fluorescent tag by exploiting the complementarity of single-stranded RNA and DNA [18]. Raj and colleagues [47] used this technique to label both transgenes and endogenous genes in mammalian cells. The authors designed a single-stranded DNA oligonucleotide to be complementary to a repetitive region of their gene of interest, and labeled each oligonucleotide with a single fluorophore. Because multiple oligonucleotide probes bound to the repeat region, the signal from that transcript appeared as a bright spot in the microscope, considerably brighter than any background signal caused by off-target oligonucleotide binding. This technique was performed in fixed cells, so it was impossible for the authors to visualize transcriptional bursting over time. However, they used an alternative method to study bursting, noting that sites of active transcription in the nucleus were characterized by bright accumulations of oligonucleotide probe. They were therefore able to identify sites of active transcription in the nucleus and record the size and frequency of these transcriptional bursts. They found that both their transgene and the RNA polymerase II gene exhibited bursty

transcription, marked by infrequent, but bright, sites of active transcription. Using the RNA FISH technique, the authors were able to not only observe sites of transcription, but also count the number of single molecules of RNA in each cell. They observed highly variable numbers of RNA molecules from cell to cell, both of the transgene and even the gene encoding a subunit of RNA polymerase. Zenklusen et al. [77] used a similar technique in yeast to show that, interestingly, endogenous genes can exhibit either bursty or Poissonian transcription.

More recently, Raj and colleagues expanded this technique to label any endogenous gene, by designing multiple individually-labeled oligonucleotides that each target a sequence on the gene of interest [49]. By designing $\sim$30 probes for a single transcript, they were able to achieve similarly high signal-to-noise ratios as they were when targeting a repeat region. Many groups have independently used this technique to show that endogenous genes display bursty transcription and that RNA abundance is highly variable from cell to cell in a population [58, 65]. This version of RNA FISH is the technique we use to measure gene expression in this thesis.

We have now seen a large body of evidence suggesting that not only is gene expression bursty in both prokaryotes and eukaryotes, but also that gene expression appears to be variable between cells. The implications of this variability for cellular homeostasis are still unclear. It is important to note that in the studies listed above, the authors were primarily comparing the absolute amount of RNAs or proteins between cells, without taking cell size into account. If RNA is expressed at a constant concentration, however, we would not necessarily expect the absolute amount to remain constant from cell to cell. It is possible that most of the observed variability is in fact due to the scaling of RNA production with volume. Furthermore, there is some evidence, detailed nicely in this review [34], that suggests that transcription kinetics may be modulated globally, perhaps by an extrinsic factor such as cellular volume.

If it is true that transcription and RNA number scales with volume at the single cell level, it will be necessary to revisit our understanding of stochastic gene expression and cell-to-cell heterogeneity. That said, there are some extreme examples of differences in gene expression between cells [47] which likely cannot be explained simply by volume. We must better understand the range of variability in gene expression and determine how, or even whether, gene expression scales with cellular volume. We have argued that, in order for cells to maintain homeostasis, they must express RNA—especially RNA from "housekeeping" genes and other genes important for cells to function properly—in concordance with volume to some extent. However, it remains to be seen how many genes express RNA in such a manner.

1.3 Modeling gene expression

If gene expression does depend strongly on cellular volume, then it will be necessary to revisit our models of gene expression "noise", which heretofore have been primarily based on variability in RNA number between cells.

The standard model of gene expression is based on the "random telegraph model" [45]. In this model, genes stochastically fluctuate between ON and OFF states. Genes can only be transcribed when the gene is ON, resulting in random bursts of transcription. According to the random telegraph model, the gene switches between ON and OFF states at exponentially distributed intervals and transcription is assumed to be a Poisson process when the gene is ON. In terms of reactions, the model is as follows:

$$I \xrightarrow{k_{on}} A$$

$$A \xrightarrow{k_{off}} I$$

$$A \xrightarrow{\mu} A + M$$

$$M \xrightarrow{\delta} 0,$$

where I indicates an inactive gene, A indicates an active gene, and M is mRNA. The model as it stands does not take volume, or any other extrinsic source of variability, into account. If overall RNA abundance does, in fact, scale with volume, then at least one of the rates, likely either the transcription rate μ or the degradation rate δ, would need to scale with volume. We discuss this modified model further in Chapter 4.

1.4 Overview

In the first half of this thesis, we use single molecule RNA imaging and computational image analysis to measure transcript abundance and cellular volume simultaneously in individual human cells, showing that, for many genes, RNA abundance does scale with volume. We show that this scaling is due to increased transcription, not reduced degradation, in larger cells. We use cell fusion experiments to show that cellular size can directly influence gene expression via a global transcriptional control, and it does so through modulation of transcriptional burst size. Furthermore, quantitative analysis of these experiments reveal that the mechanism underlying this global regulation does not merely sense cellular volume, but rather integrates both DNA content and cellular volume to produce the appropriate amount of RNA for a cell of a given size. We also show that a separate mechanism exists to reduce the frequency of transcription immediately upon DNA replication, which prevents early-replicating genes from producing an excess of RNA during early S phase.

The observation that RNA expression scales with cellular volume led us to reconsider the traditional measures of "noise" in gene expression. In collaboration with our

colleague Abhyudai Singh, we developed a quantitative framework for interpreting gene expression variability in single human cells, which we discuss in Chapter 4. We show that many genes that would traditionally be considered noisy actually display noise levels close to Poisson when volume dependence is taken into account. To further extend this noise analysis, we performed single-cell RNA sequencing and calculated noise measures genome wide, revealing that cell-type specific genes are more variable than ubiquitously expressed genes.

Some of the work presented here has appeared in print in the following publications and has been reprinted with permission:

- O. Padovan-Merhar et al., *Molecular Cell* (Set for publication on April 9, 2015).
- M. N. Cabili et al., *Genome Biology* **16** 20 (2015).

Chapter 2

What's really going on here? Understanding the scaling of biomolecules with cellular volume

2.1 We measure mRNA counts and volume in single cells using RNA FISH

We first looked at the number of mRNA molecules in individual primary fibroblast cells (human primary foreskin fibroblasts, CRL2097) within a population to see whether mRNA counts scale with cellular volume at the single cell level. We measured both mRNA abundance and volume simultaneously using single molecule multi-color mRNA fluorescence *in situ* hybridization (RNA FISH [18, 49]), which allowed us to detect the positions of individual mRNAs in three dimensions as fluorescent spots in the microscope (Fig. 2.1).

With the filter sets on our microscope, we can detect up to six colors simultaneously. For most experiments, we labeled (1) the mRNA from our gene of interest, (2) the introns from our gene of interest, which allows us to accurately detect sites of active transcription, (3) *GAPDH* mRNA, which is highly expressed and allows us to calculate

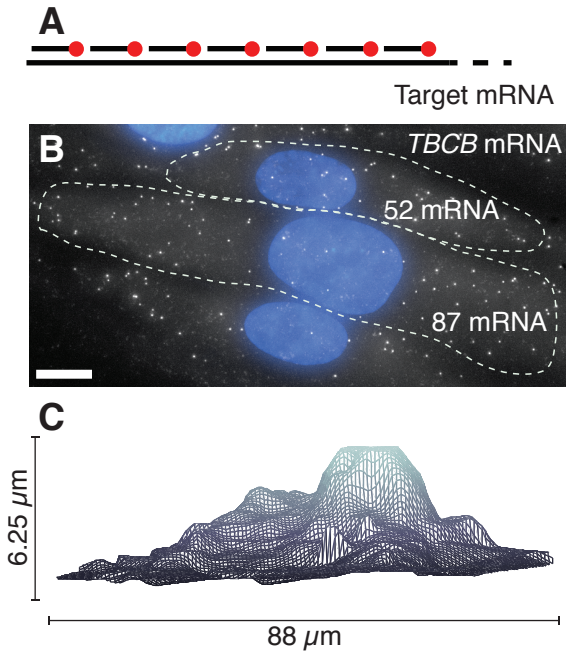

Figure 2.1: Representative RNA FISH image and volume measurement. (A) Schematic of RNA FISH technique. We designed 20-base oligonucleotide DNA probes with sequences complimentary to our target mRNA. We designed ∼30 probes, each labeled with an individual fluorophore, to tag a single species of mRNA, giving us a high signal-to-noise ratio. (B) Single molecule RNA FISH. DAPI stain in blue, *TBCB* mRNA FISH probe in white. Scale bar is 10μm. (C) Representative outline of a primary fibroblast cell found using our volume calculation algorithm.

the 3D boundaries and volume of the cell, (4) *CCNA2* mRNA, a cell cycle marker, (5) actin protein, which allows us to accurately visualize the boundaries of the cell in 2D, and (6) DNA with DAPI to visualize the nucleus.

We measured mRNA abundance and volume in single primary fibroblast cells for 30 different genes, calculating volume as described in the Methods (Appendix A). Briefly, we detected the 3D locations of all *GAPDH* mRNA molecules in the cell, which fill the volume of the cell. We computationally identified the outermost molecules and interpolated those points to define the top and bottom boundaries of the cell, and we calculated volume by summing the height difference between top and bottom. Note that this method will systematically underestimate the volume of the cell, so we use a statistical algorithm to correct for bias in volume measurement. We show a representative RNA FISH image and cellular volume schematic in Fig. 2.1. Note that the fibroblast cells are adherent and have a characteristic "fried egg" shape, making it difficult to estimate volume in any way other than actually determining the boundaries through RNA FISH. We repeated these measurements for 25 genes in a lung cancer cell line (A549) and for 19 genes in growth-arrested fibroblast cells (see Appendix B).

The cell volumes we measured varied over an approximately six-fold range, ranging on average from 1 to 6 picoliters (pL), although the largest fibroblast we measured had a volume of 8.95pL, and the smallest had a volume of 0.443pL. These measurements agree with measures of mammalian cellular volume obtained using different methods. Bryan et al. [7] measured volumes of lung cancer cells using both a Coulter counter and a microfluidic device, a "suspended microchannel resonator", and found that these cells had a mean volume of ∼3pL, with a range from 1-7pL. Tzur et al. [67] measured volumes of mouse lymphoblasts using a Coulter counter, and found the cells to have volumes of ∼1.5pL, ranging from ∼0.5-3pL. Zhao et al. [78] estimated HeLa cell volumes to be approximately 2.6pL by assuming them to be spherical and

extrapolating from their diameter as viewed in a light microscope. Our measurements of cell volume have a similar mean and variance to those reported in these studies, so we believe our method of calculating volume is accurate.

To test that our measurement was not biased by our choice of guide gene, we calculated volume using both *GAPDH* (RNA count = 2762±156.7, volume = 2438±170.3pL) and *EEF2* (RNA count = 1063±63.51, volume = 2147±160.2pL), finding that both methods yielded similar results (Fig. 2.2). We also calculated volume by randomly selecting only half the number of *GAPDH* mRNA spots, and found that volume was the same before and after reducing the number of spots (Fig. 2.2). Thus we have shown that our volume calculation is robust to fluctuations in the total number of *GAPDH* mRNA and that our metric is not biased by our choice of guide gene.

All of our measurements were taken in fixed cells, which had been treated with formaldehyde to cross-link and preserve distances between all molecules within the cell. It is possible that the fixation procedure could systematically change cellular volume, so we therefore calculated "volume" as best we could throughout the fixation process. We introduced fluorescent beads to the media of live cells, and allowed the beads to settle onto the tops of the cells. We then imaged the cells in 3D, identifying the top of the cells by the location of the beads, and the area of the cells through brightfield images. Without removing the cells from the microscope, we added formaldehyde to the cells and allowed them to "fix" as usual. After fixation, we repeated the measurement with the fluorescent beads. We finally added ethanol to permeabilize the cells, and repeated the measurement a final time. Heights and areas of the cells are shown in (Fig. 2.2), and qualitatively agree with the measurements of volume that we calculate through our RNA FISH method. Further, the measurements for each cell did not change significantly or systematically throughout the fixation and permeabilization procedure, so we have shown that there are no large biases in volume measurements

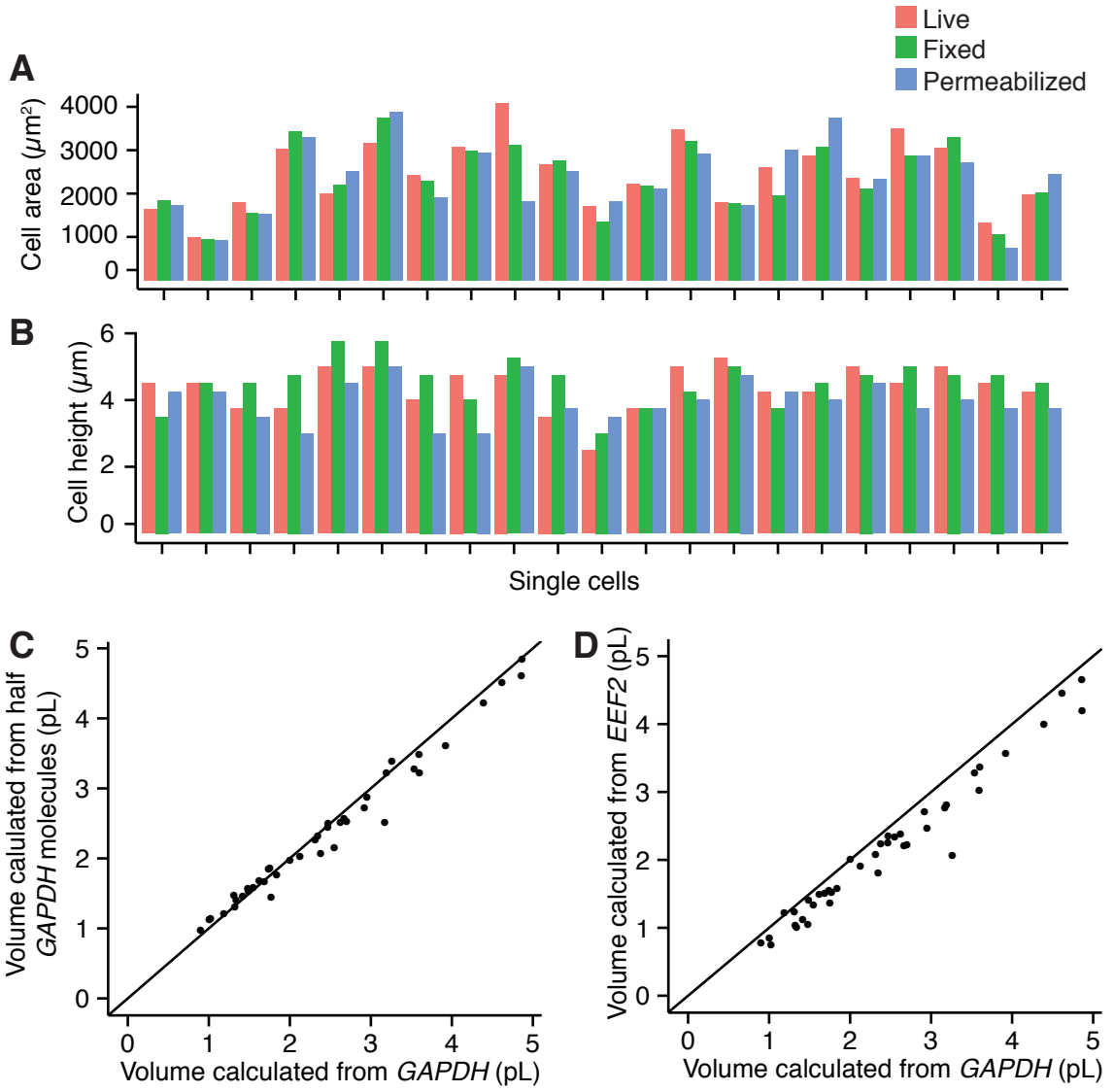

Figure 2.2: Volume calculation controls. (A) and (B) We monitored cells on the microscope throughout the process of fixation. We took measurements of the same cells live, after fixing in 4% formaldehyde for 10 minutes, and after permeabilizing in 70% ethanol for 30 minutes. (A) We measured the areas of the cells through brightfield images. (B) We measured the height of the cells by coating the cells with fluorescent beads. (C) To demonstrate the robustness of the volume calculation algorithm, we calculated volume for the same cells using all the *GAPDH* mRNA spot coordinates as detected by RNA FISH, or using only half of the points, chosen randomly. (D) We calculated volume using a different gene, *EEF2*. Black lines indicate a fit with intercept = 0 and slope = 1.

that arise from fixation. We also note that all cells behaved similarly during this process, so fixation does not lead to random variability in volume between cells.

2.2 mRNA counts scale with cellular volume in single mammalian cells

We measured RNA and volume in single cells using the methods described above. For most genes, mRNA counts and volumes in single cells exhibited a strongly positive, linear correlation (we show data from a few representative genes in Fig. 2.3), although mRNA from many genes—primarily transcription factors and other "non-housekeeping" genes—displayed less of a correlation, which we discuss further in Chapter 4. See Appendix B for all genes examined. Because in general larger cells had proportionally more transcripts than smaller cells, the mRNA concentration remained relatively constant from cell to cell despite considerable variation in absolute mRNA numbers. This scaling property was not confined to high abundance mRNAs like *GAPDH* and *EEF2*—genes expressing as few as 10-20 mRNA per cell such as *ZNF444* and *KDM5A* scaled similarly. We also measured rRNA, by far the most abundant RNA in the cell. rRNA is so abundant that it is impossible to resolve individual molecules by RNA FISH, so we instead summed the total intensity of the RNA FISH signal. We found that this quantity scaled with cellular volume as well (Fig. 2.4), suggesting that RNA products from both RNA polymerase I (rRNA) and II (mRNA) display similar scaling properties. We also observed the same behavior for short lived mRNA such as *UBC* and *IER2*, whose half-lives are 2.9 and 2.2 hours, respectively [62]. These results show that this scaling property is universal, and not simply limited to highly abundant or long-lived transcripts.

It is possible that the question of how RNA scales with volume is simply answered

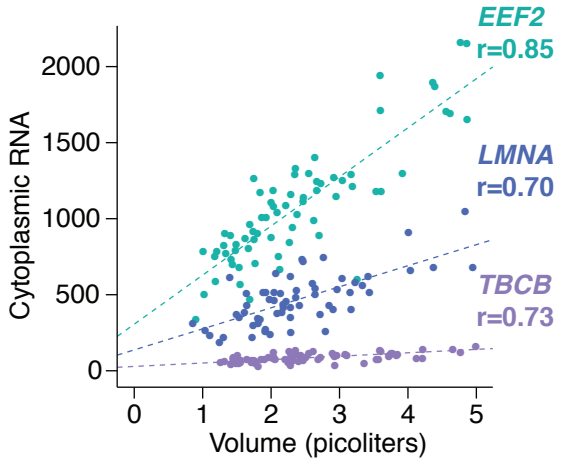

Figure 2.3: RNA count scales with volume for many genes. mRNA vs. volume for *EEF2*, *LMNA*, *TBCB*. Each point represents one single cell measurement. Each data set is a combination of at least two biological replicates, with at least 30 cells per replicate. Dashed lines represent best fit via linear regression.

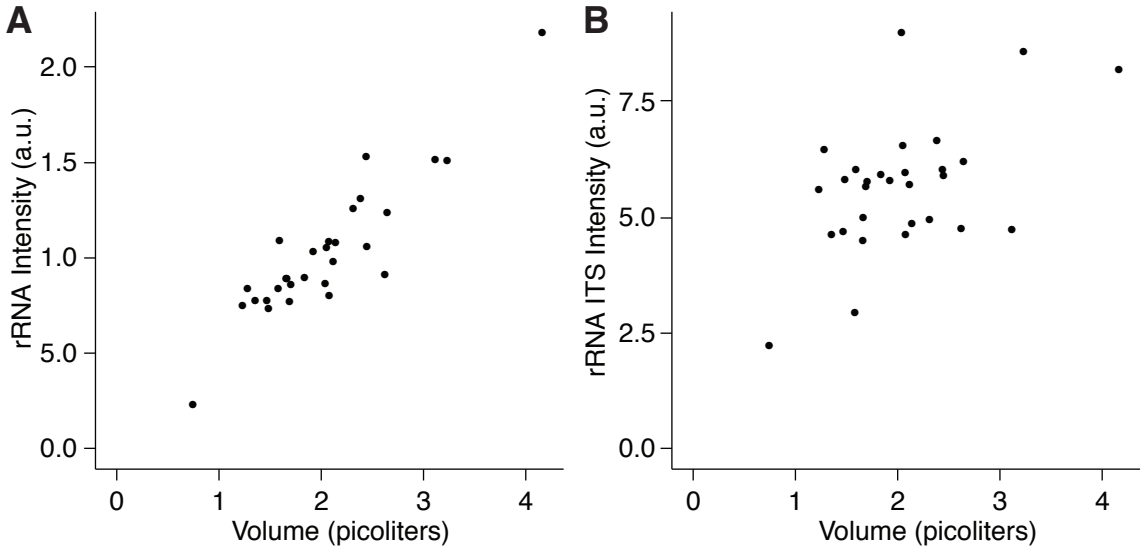

Figure 2.4: Ribosomal RNA. (A) We measured ribosomal RNA by quantifying total fluorescence intensity in the cytoplasm from an rRNA FISH probe in cycling fibroblast cells. (B) We measured the rRNA internally transcribed spacer (ITS, the rRNA "intron") by quantifying total fluorescence intensity in the nucleus from an ITS RNA FISH probe.

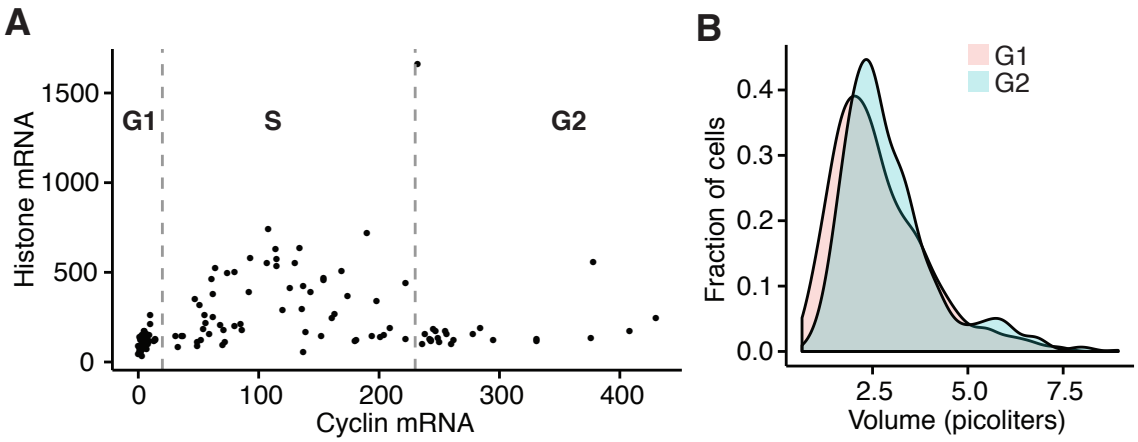

Figure 2.5: Cell cycle determination. (A) We simultaneously measured *CCNA2* and *HIST1H4E* mRNA by RNA FISH to precisely determine cell cycle position. Each data point is a single cell measurement. Data shown are from one of four biological replicates. (B) Volume distributions in G1 and G2. We determined cell cycle position using *CCNA2*. n = 841 cells in G1, 191 cells in G2.

by considering cell cycle progression. As a cell progresses through the cell cycle, it grows (theoretically doubling in size) and also replicates its DNA. One might expect that the larger cells we observe are simply further along in the cell cycle and have already replicated their DNA. In this scenario, larger cells have twice the number of DNA molecules as compared to smaller cells, and therefore can produce more RNA. To address this possibility, we co-stained cells with cell cycle markers [17, 29, 53, 73] to differentiate between G1, S, and G2 phases. Cyclin A2 (*CCNA2*) mRNA is expressed only in S and G2 phases, while Histone 1, H4e (*HIST1H4E*) mRNA is expressed only in S phase. We classified cells as being in G1 if both *CCNA2* and *HIST1H4E* mRNA levels were low (cutoff = 20 *CCNA2* mRNA), S if *CCNA2* was mid-range and *HIST1H4E* was high, and G2 if *CCNA2* was high and *HIST1H4E* was low (cutoff = 230 *CCNA2* mRNA) (Fig. 2.5).

Interestingly, cell volume varied as much for cells in individual phases of the cell cycle as the population overall, with a shift in the distribution towards G2 cells being

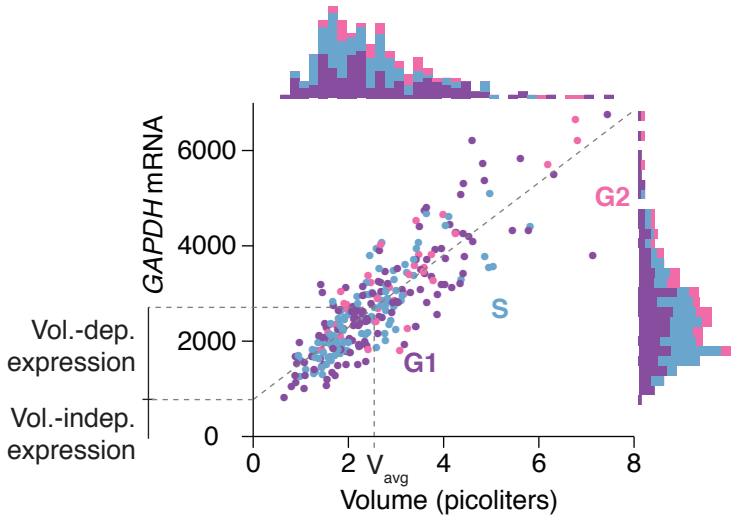

Figure 2.6: *GAPDH* RNA scales with volume similarly throughout the cell cycle. Marginal histograms show volume and mRNA distributions. Colors indicate cell cycle stage determined by Cyclin A2 (*CCNA2*) mRNA count. Dashed diagonal line is the best linear fit of RNA vs. volume. V_{avg} indicates the average primary fibroblast volume. We determined volume-independent and -dependent transcript levels using the linear fit and V_{avg}. Data are a 15% subset of 1868 cells spanning >30 biological replicates.

larger, but only by ∼10% (Fig. 2.5). Although it might initially be surprising that G2 cells are not on average twice as large as G1 cells, we believe this can be explained by non-uniform growth throughout the cell cycle. If, for an extreme example, cells remained the same size throughout the cell cycle, only doubling in size immediately before division, we would not expect a population average of cells to show that cells in G2 are twice the size of those in G1. The size differences that we observe are likely due to a less extreme case of non-uniform growth during the cell cycle. Further, we observed the same linear relationship between mRNA and volume for cells restricted to a particular phase of the cell cycle as for all cells together (Fig. 2.6), showing that mRNA count did not depend on DNA content of the cell.

In addition to cellular volume, we also measured nuclear size throughout the cell cycle. Our imaging techniques did not allow for a 3D nuclear volume measurement,

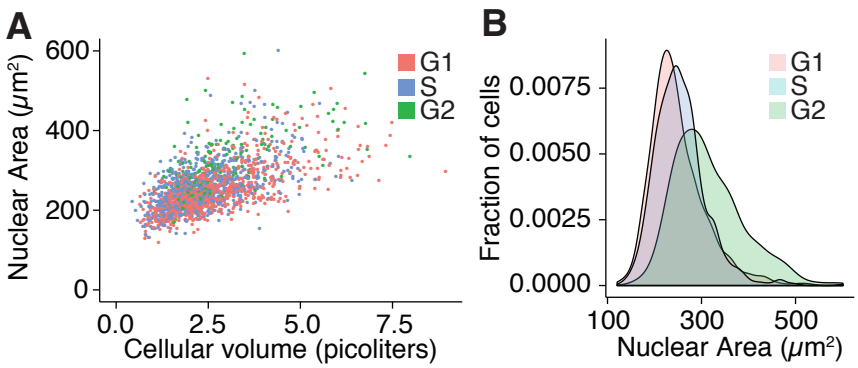

Figure 2.7: Nuclear area scales with cell cycle stage and cell size. (A) Nuclear area vs. cytoplasmic volume. We measured cytoplasmic volume using our standard method. We measured nuclear area using the DAPI stain. We note that we only measured nuclear area and not volume. $R^2 = 0.358$. (B) Density plot of nuclear area across cell cycle stages. n = 1866 cells.

so we approximated nuclear volume by nuclear area, defined by the DAPI stain. We found that nuclear size increased somewhat with cellular volume, and that nuclear size increased in later stages of the cell cycle (Fig. 2.7). It is interesting that different cells in the same stage of the cell cycle display different nuclear sizes. Such cells have the same number of DNA molecules, so a larger nucleus implies that the same amount of DNA is taking up more space. This larger nucleus could be the result of more active transcription, as it is likely that DNA bulges and takes up more space during active transcription [64]. It could also be the effect of simply having a larger nuclear envelope. In this case, everything in the nucleus is just less dense. We explore the consequences of different nuclear sizes in Section 3.3.

It has been shown in yeast and bacteria that cell size scales with ploidy [38, 40, 66]. It could be the case that in our system, larger cells have a higher ploidy, leading to a larger cell volume and increased RNA abundance. However, we note that the primary fibroblast cells exhibited normal ploidy [29], so our results are not simply explained by differences in ploidy between cells. Additionally, the A549 lung cancer cell line is

known to have abnormal ploidy, containing more than 2 copies of some genes [23]. Despite the abnormal number of DNA molecules, we still observe that RNA scales with volume in these cells, demonstrating that cells have a means of maintaining volume scaling despite differences in DNA content. Together, these results show that ploidy differences do not change our qualitative observations, and that our results are not simply due to differences in ploidy among cells.

We next wanted to check whether mRNA scaling is dependent upon continual cellular growth, or if the scaling continues even in the absence of active growth. We grew the primary fibroblasts for 7 days in medium lacking serum, making them quiescent. The cells ceased proliferating and were all arrested in G0, as indicated by cell cycle markers. Despite growth and cell-cycle arrest, we found that mRNA count and volume still scaled strongly, showing that neither progression through the cell cycle nor continual cell growth were required for mRNA count to scale with cellular volume. Interestingly, we found that both the mean mRNA count and mean volume decreased in quiescent cells, although the cells maintained a similar concentration of *GAPDH* and other mRNA between the two conditions (Fig. 2.8). This finding has an important implication for single-cell studies. By simply counting mRNA between the two conditions, we likely would have come to the conclusion that there is a fundamental difference in activity or cellular function between the two conditions, as the quiescent cells have less mRNA overall. However, by taking volume into account, we saw that both conditions actually had the same mRNA concentration and likely had similar "activity" levels, despite the lack of growth in the quiescent cells. This finding highlights the importance of including volume as a variable in such single-cell measurements.

It is important to note that while the mRNA abundance is strongly correlated with cellular volume, the y-intercept (a) of a line fit to the data ($mRNA = a + bV$) was non-zero, indicating that mRNA count in individual cells had a volume-independent

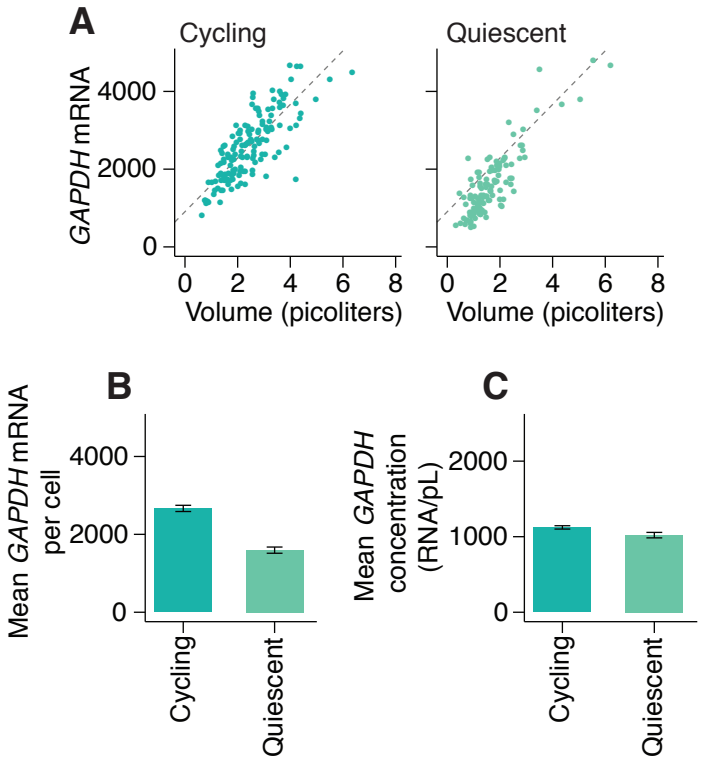

Figure 2.8: Comparison of *GAPDH* RNA expression in cycling and quiescent primary fibroblast cells. (A) Dashed lines are best fit line for *GAPDH* in cycling cells. Data are an 8% subset of 1868 cells spanning >30 biological replicates for cycling cells, and 10% subset of 1105 cells for quiescent. We only analyzed quiescent cells that had fewer than 20 *CCNA2* mRNAs. (B) Mean *GAPDH* mRNA count and (C) concentration in different growth conditions for data from (A).

component in addition to the volume-correlated component (Fig. 2.6). We quantified for each gene the relative fraction of mRNA that was volume-correlated vs. volume-independent in a cell of average volume (i.e., $a/(a + bV_{avg})$ vs. $bV_{avg}/(a + bV_{avg})$). We found that different genes displayed a range of values for volume-independent and volume-correlated abundance (Fig. 2.9), although the volume-dependent fraction was dominant for most genes examined. Of note, these results show that the mRNA concentration is actually somewhat greater in smaller cells than in larger ones; for most genes, the smallest cells had an mRNA concentration 1.2-3 times greater than that of the largest cells (Appendix B). We later describe a mathematical model providing a potential explanation for this increased concentration based on nuclear volume measurements (see Section 3.3).

We compared volume-correlated and -independent fractions in cycling and quiescent cells to see if there was a link between growth and volume-correlated or -independent mRNA abundance. Interestingly, although the overall mRNA concentration was similar between conditions, both volume-correlated and -independent abundance was lower in quiescent (growth-arrested) cells, although there was a more significant difference between the two conditions for the volume-independent abundance (Fig. 2.9). This suggests that there may be a link between volume-independent expression and growth. Perhaps cells produce mRNA in a volume-independent manner to trigger or continue growth. Our data is not sufficient to make a substantial claim about this phenomenon, but it is an interesting hypothesis.

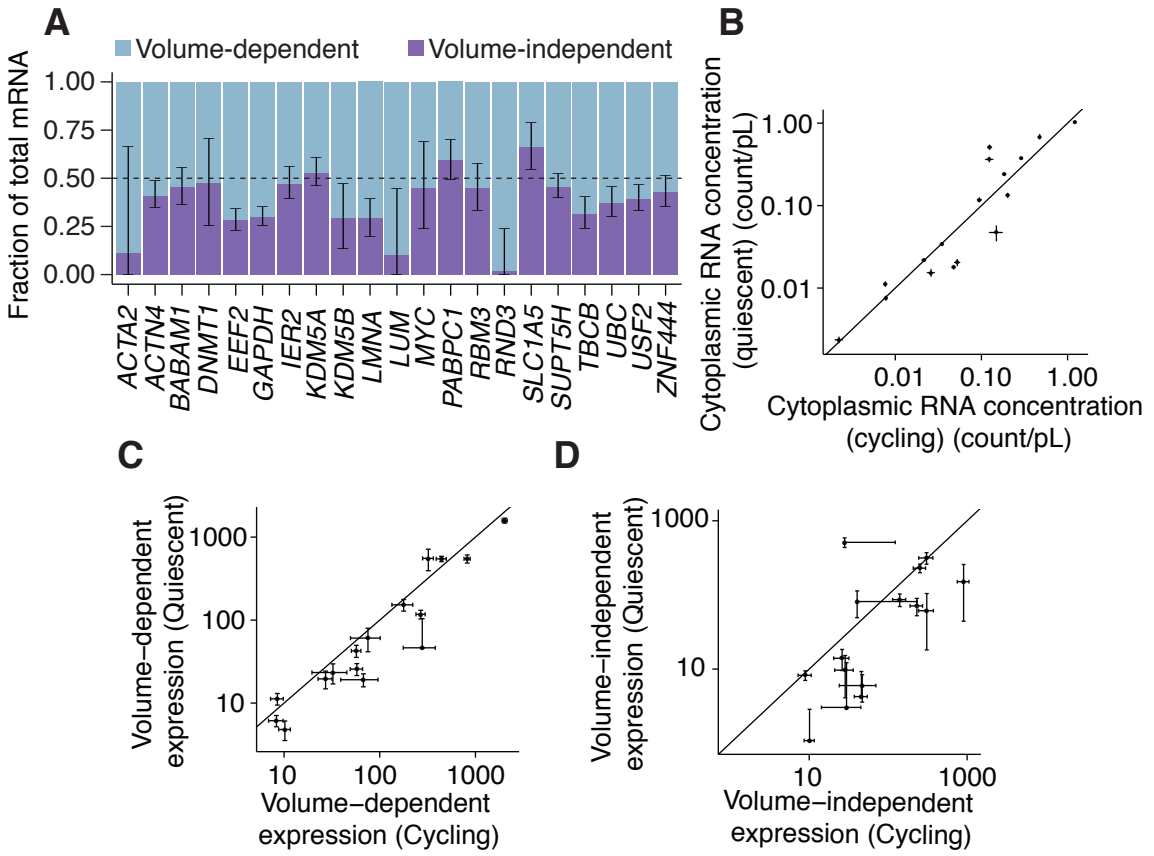

Figure 2.9: Volume-dependent and volume-correlated RNA abundance in fibroblast cells. (A) Fraction of volume-independent and -dependent RNA expression from the linear fit of RNA vs. volume for 21 genes in primary fibroblast cells (we omitted highly variable genes whose volume-independent fractions were less than zero). Data for each gene are a combination of at least two biological replicates, with at least 30 cells per replicate. (B) Concentration of mRNA in cycling and quiescent fibroblast cells. Each data point represents a single gene. Error bars represent standard error of the mean. (C) and (D) We compared volume-dependent and -independent abundance for cycling and quiescent cells. All error bars represent confidence intervals of the slope or intercept of the fit, normalized to the scale of the plot. We omitted error bars that extended below zero. Each gene had a minimum of two biological replicates, with at least 30 cells per replicate. We omitted highly variable genes with intercept terms less than 0.

2.3 RNA counts scale with organism size in *C. elegans*

We have shown that mRNA count scales with volume in single cells, and we wanted to check whether we could observe similar behavior in intact organisms. *C. elegans* is a small nematode, and is a model organism touted for its optical transparency and genetic tractability. Because these worms are transparent we are able to perform RNA FISH on intact organisms. Thanks to the tractable genetics of *C. elegans*, we obtained two previously-generated strains of worm from the Caenorhabditis Genetics Center that allowed us to look at worms with cells of different sizes. CB502 worms have a mutation in the gene *sma-2*, leading to a small body size; N2 worms are wild-type and have a body size approximately twice that of CB502 (Fig. 2.10). Despite these differences in size, these two types of worm have the same number of cells. In general, each cell is simply smaller in CB502 worms, although some of the intestinal cells also have different ploidy [71].

Each nematode was too large to image completely, so we chose regions of the worm that were easily identifiable between specimens, namely the head and gonad regions. We measured both RNA and DNA density in the heads and gonads of adult nematodes, comparing measurements from both N2 and CB502 worms. We measured mRNA expression of two different genes: *ama-1*, encoding a subunit of RNA polymerase II, and *arf-3*, encoding an ADP ribosylation factor. Both genes have "housekeeping" functions, so we predicted that their mRNA would scale with volume. We measured concentration of these mRNAs by performing RNA FISH and counting spots as usual, then normalizing to the volume of the imaged area, approximated simply by a rectangular prism defined by the field of view. We found that the RNA concentrations were roughly the same between the two strains, despite the smaller strain having

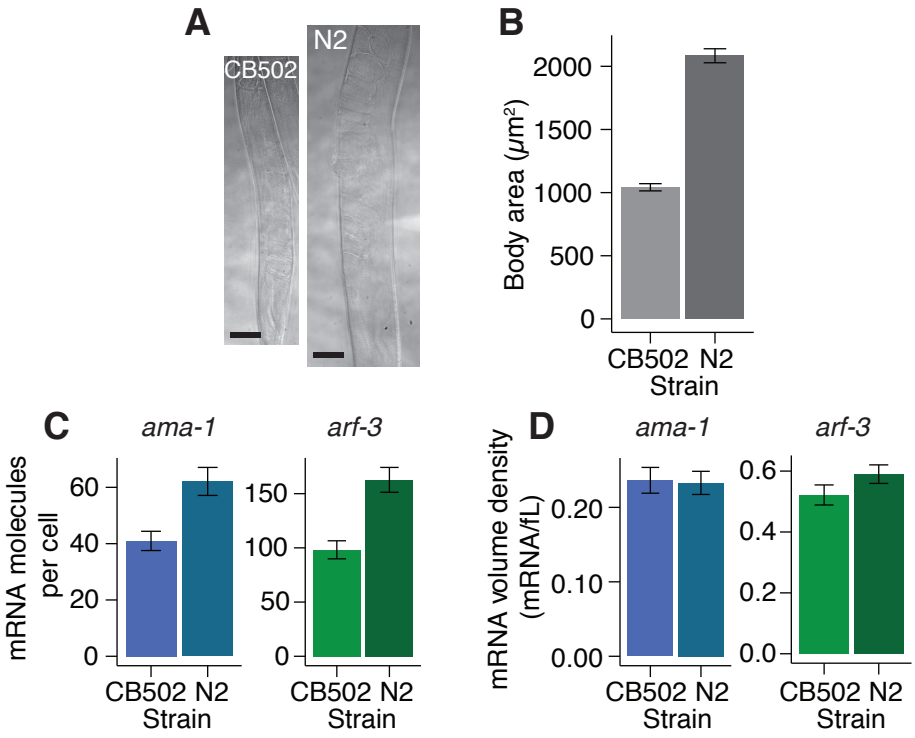

Figure 2.10: RNA scales with volume in *C. elegans*. (A) Images of the two *C. elegans* strains. Scale bars are 10μm. (B) Sizes of the two strains. (C) Number of mRNA molecules per cell in the gonad region for each type of worm for genes *ama-1* and *arf-3*. We estimated the number of cells in each segment by counting nuclei stained with DAPI. Each bar is a compilation of 3 biological replicates, with >3 worms per replicate. (D) Concentration of mRNA in the gonad region. All error bars represent standard error of the mean.

smaller cells overall (Fig. 2.10). It is difficult to distinguish single cells within an intact organism, but we approximated the number of RNA per cell by normalizing the total amount of RNA in a field of view by the number of nuclei in the same field. As expected, the number of RNA per cell decreased by a factor similar to that of the volume differences between the strains, verifying that our observations can hold in vivo.

2.4 Transcriptional activity, not mRNA degradation, scales globally with cellular volume

We have shown that larger cells have a proportionally higher number of mRNA than smaller cells, even if they have the same absolute number of DNA molecules. To maintain this proportionality, larger cells must either transcribe more mRNA from the same number of DNA molecules or degrade those mRNA more slowly. Mathematically:

$$\frac{dm}{dt} = \mu(V) - \delta_0 m \tag{2.1}$$

or

$$\frac{dm}{dt} = \mu_0 - \delta(V)m, \tag{2.2}$$

where m is the number of mRNA molecules, μ is the transcription rate, and δ is the degradation rate. Assuming that only either μ or δ is volume-dependent, to achieve the scaling that we observe, we likely have $\mu(V) = \mu_0 V$ or $\delta(V) = \delta_0/V$, although it is possible that both production and degradation terms have some volume dependence (see below). To distinguish between the two possibilities above, we determined the rate of mRNA degradation in cells of different sizes. We treated cycling primary fibroblast cells with actinomycin D, a drug that inhibits transcription by binding to the transcription initiation complex and inhibiting elongation. We performed RNA FISH on two populations of cells: one was untreated, and the other was treated with 100nM actinomycin D for four hours. Since the treated cells had no active transcription, all observed mRNA dynamics were the result of degradation, and we assumed that the data could be fit by the following functional form:

$$m(t, V) = m_0(V)e^{-t/\delta(V)}. \tag{2.3}$$

Indeed, we observed exponential decay over the course of a few hours (Fig. 2.11). Comparing the untreated to the treated cells, we used the fit line for the untreated cells to determine $m_0(V)$, and determined $\delta(V)$ for every treated cell we measured. We found that degradation rate did not change significantly in cells of different volumes (Fig 2.11), and indeed the data was much better fit by the model in which $\delta(V) = \delta_0$ than by the model in which $\delta(V) = \delta_0/V$. We therefore have shown that that reduced degradation is not responsible for the increased number of mRNA in larger cells.

We also performed a second analysis in which we fit the data from the transcriptionally inhibited cells to a discriminatory model with a fitting parameter α:

$$m(t, V) = m_0(V)e^{-tV^\alpha/\delta_0}. \tag{2.4}$$

A value of $\alpha = 0$ indicated that the data was best fit by a model in which degradation rate (δ) was independent of volume. A value of $\alpha = 1$ indicated that the data was best fit by the model in which $\delta = \delta_0/V$. An intermediate value of α suggested the possibility that both μ and δ had some volume dependence. As expected given our previous data, we found that, for these genes, the best-fit model had a value of α that was within error of 0, therefore showing in a second way that degradation happens in a volume-independent manner.

We next checked whether larger cells transcribe more than smaller cells (as observed in bulk populations [19, 55, 79]). We inferred global transcription rate by incorporating a labeled uridine into all newly synthesized RNA produced during a 60 minute time window (Fig. 2.12), which we then rendered fluorescent via click chemistry [24]. Note that mRNA with the labeled uridine had a nuclear export defect, so most of the labeled nascent mRNA was sequestered in the nucleus. Therefore, the total intensity of fluorescence within the nucleus was proportional to the total amount of

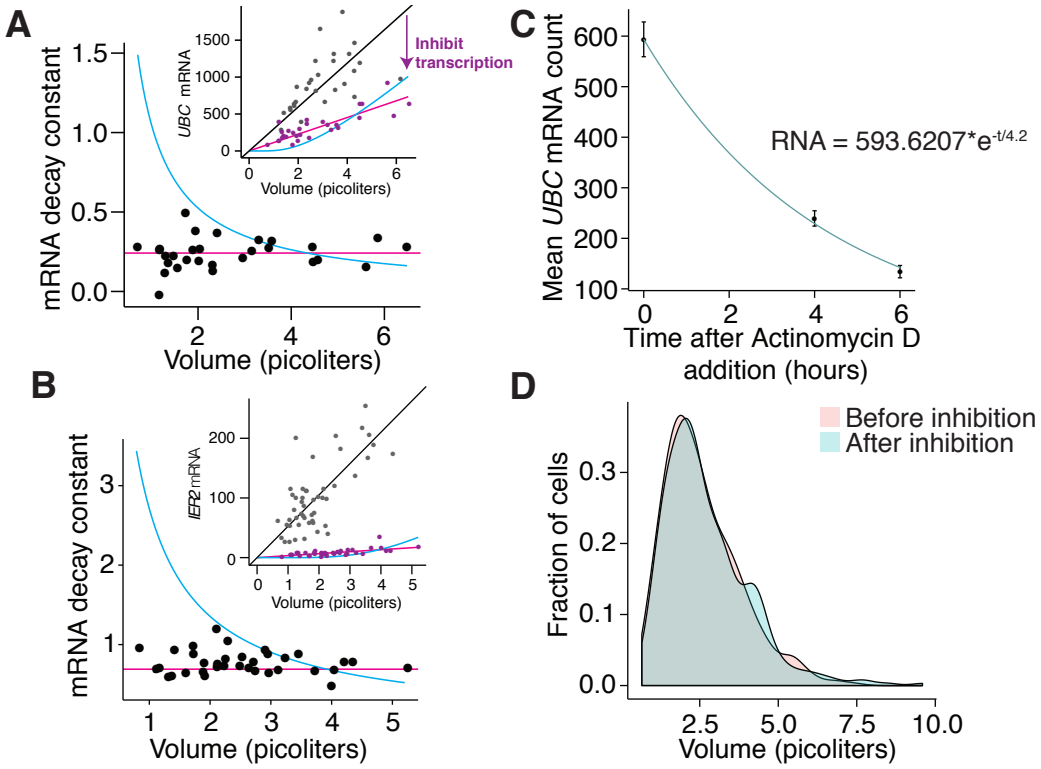

Figure 2.11: Degradation rate is independent of volume. (A) and (B) We inhibited transcription in primary fibroblast cells using actinomycin D for 4 hours and allowed *UBC* (A) or *IER2* (B) mRNA to degrade. Inset shows mRNA before and after inhibition. Each point represents a single-cell measurement. We calculated the decay constant for each cell using the best-fit line before inhibition (see Appendix A). Blue line shows fit if degradation were volume-dependent; red line shows fit if transcription were volume-dependent. Data represent one of two biological replicates. (C) RNA degrades exponentially when transcription is inhibited with actinomycin D. Pictured is *UBC* mRNA 0, 4, and 6 hours after transcription block. (D) Distribution of cell volumes before and after inhibition by Actinomycin D. The volume distribution is similar before and after we inhibit transcription. Although we cannot track a single cell before and after inhibition, this suggests that actinomycin D likely does not change the volume of a cell, so it is appropriate to use the fit line before inhibition to calculate the decay constant. n = 459 cells before inhibition, 413 cells after inhibition.

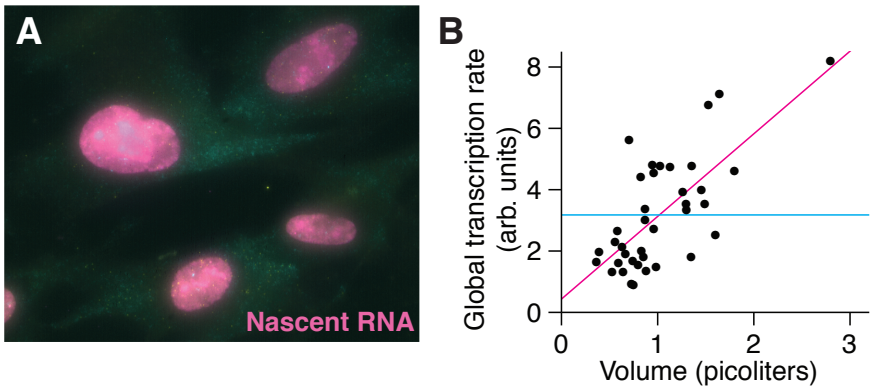

Figure 2.12: Transcription rate correlates with volume. (A) We fluorescently labeled nascent RNA produced in one hour using the Click-iT eU assay in primary fibroblast cells. Pictured is raw micrograph data. (B) We quantified the total fluorescence intensity (transcription rate) by imaging the nuclei of single cells. Blue line shows fit for volume-dependent degradation; red line shows fit for volume-dependent transcription. Data shown is from quiescent cells, and is one of three biological replicates.

new transcription during the 60 minute time window. The total nuclear fluorescence was therefore equivalent to the global transcription rate. We found that transcription rate was linearly proportional to volume, thus showing that individual cells vary in their overall transcription [13] and these variations correlate strongly with volume. das Neves et al. [13] performed a similar transcription rate assay, and also found that transcription rate was variable between cells, even suggesting that transcription rate may depend on a small diffusible factor. However, the authors connected transcription rate to the mitochondrial content of single cells, not cellular volume. It may be the case that both volume and mitochondria play a role in regulating transcription. From our data, however, we conclude that volume is linked to transcription rate and that larger cells maintain proportionally higher levels of RNA by increased transcription rather than decreased degradation as compared to smaller cells.

We also quantified fluorescence from probes targeting the internal transcribed

spacer of the rRNA (the "intronic" sequence of rRNA) and showed that transcription of rRNA scaled linearly with volume (Fig. 2.4), indicating that RNA polymerase I transcription is also volume-dependent.

2.5 A global mechanism links transcription and volume

The scaling of transcription with cellular volume could be due to (1) global factors regulating transcription of all genes in a volume-correlated manner, or it could be that (2) gene regulatory networks sense deviations in each particular gene's protein concentration and modulate transcription to restore concentration. In the latter case, reducing protein concentration of any one gene would result in increased transcription to compensate, whereas in the global scenario, reducing the concentration of any one gene product would not appreciably affect the cellular volume, thus leaving the gene's transcription unchanged. We tested this by reducing the level of Lamin A/C mRNA and protein in the cell via small interfering RNA (siRNA) (Fig. 2.13); we chose Lamin A/C because its expression scales strongly with volume (Fig. 2.3) and is thought to be tightly regulated [61].

To measure the transcriptional response, we took advantage of the fact that transcription occurs in bursts [8, 12, 20, 47, 59, 69, 77]. Each gene on the DNA can be in an "ON" or an 'OFF" state, and only when a gene is ON does it produce RNA. Different genes have different bursting kinetics, and we can define transcriptional activity in terms of "burst size"—the number of RNA produced during a single ON state—and "burst frequency"—how frequently the gene is in an ON state.

Genes that are actively undergoing a transcriptional burst have bright accumulations of nascent RNA at the site of transcription itself [29, 30, 47, 77]. Note that because

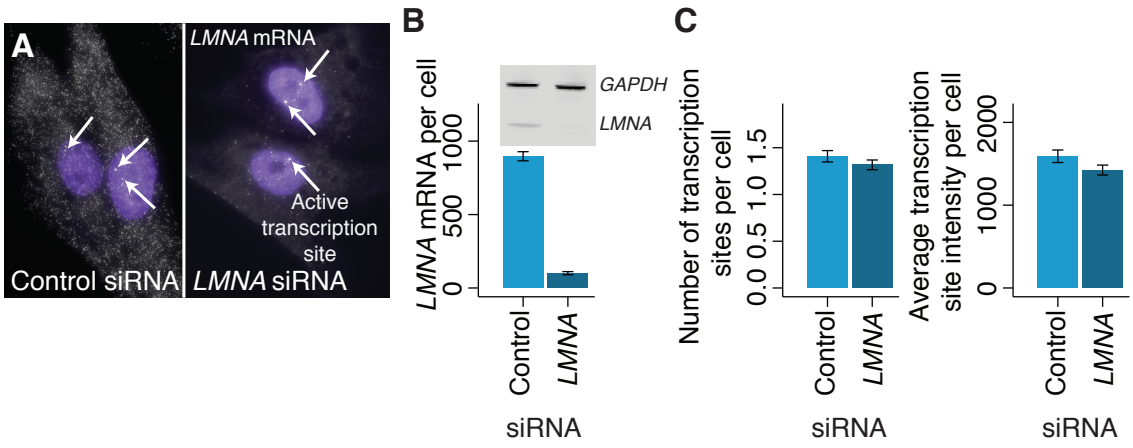

Figure 2.13: Transcription remains the same after protein knockdown. (A) We performed siRNA treatment for 72 hours in primary fibroblast cells using either a control siRNA (left), or an siRNA targeting *LMNA* mRNA (right). DAPI stain is shown in purple, and *LMNA* mRNA FISH probe is shown in white. White arrows indicate active transcription sites. (B) Quantification of cytoplasmic *LMNA* mRNA knockdown by RNA FISH. Inset shows protein knockdown. (C) Comparison of the number of *LMNA* transcription sites and transcription site intensity in siRNA control and *LMNA* knockdown conditions. We detected transcription sites through intron/exon colocalization using RNA FISH. All error bars represent standard error of the mean. Data in B, C are a combination of two biological replicates, n = 323 cells for control siRNA, 284 cells for *LMNA* siRNA.

all of our measurements were in fixed cells, we could not observe genes bursting in time, so we instead relied on population measurements, as opposed to time measurements. Thus, in place of burst size, we measured the intensity of the accumulation of RNA at the transcription site, and in place of burst frequency, we measured the number of ON transcription sites in a single cell divided by the total number of DNA copies in the cell (2 in G1, 4 in G2). We term this quantity "burst fraction" to differentiate it from burst frequency. Burst fraction measures the total fraction of gene copies that are ON at a given time, which is not precisely equivalent to the frequency at which a single copy of the gene turns ON, but is nonetheless a measurement of how active the gene is in time [29]. To ensure that we were identifying transcription sites correctly, we labeled both the exons and the introns of our gene of interest, using two distinct fluorophores. Introns are spliced out at the site of transcription, generally allowing for precise detection of active transcription sites. We developed a graphical user interface (GUI) that overlays images from the exon channel and intron channel, allowing the user to manually select transcription sites. There are a few genes for which introns do not appear to be immediately spliced out, but rather diffuse throughout the nucleus still attached to the exons. Cases like this necessitate the manual identification of transcription sites.

To reduce the amount of Lamin A/C protein in the cell, we used an siRNA, which does not affect nuclear RNA [33]. We measured both the average number of active transcription sites per cell and the intensity of those transcription sites, finding both metrics unchanged upon reduction of Lamin A/C protein levels (Fig. 2.13). Because we saw that transcription was unchanged even after the protein levels were reduced, we concluded that increased mRNA counts in larger cells result from a global difference in transcription rather than the activity of a particular gene network that regulates the concentration of Lamin A/C. However, our data do not exclude the possibility

that there may be other situations in which mRNA levels are regulated by specific networks.

Chapter 3

How does it work? A mechanistic view of the cell's transcriptional compensation for size and DNA content

3.1 A diffusible *trans* factor sensing DNA content and volume links cellular volume and transcription

Thus far, we have established that there exists a strong correlation between cellular volume and the amount of mRNA in the cell. However, we are interested in establishing directionality of this correlation. It could be the case that either the total cellular content exerts a global influence on transcription, thus making transcription scale with cellular volume; or, alternatively, transcription itself may affect cellular volume. The way to conclusively distinguish between these possibilities would be to add cellular volume to a small cell and observe if and how transcription changes in response. Unfortunately, there is no straightforward way to simply add volume to cells, so we

used an alternative method.

We made cell fusions (heterokaryons [46]) combining small human melanoma cells expressing GFP mRNA (WM983b-GFP-NLS) with larger primary fibroblasts that did not express GFP (Fig. 3.1). Note that the melanoma cells did not express the *GAS6* gene, so we were able to identify heterokaryons as cells with two nuclei that expressed both GFP and *GAS6* mRNA. GFP mRNA in the melanoma cells exhibited a strong correlation with cellular volume before fusion. Because only the smaller of the two cell types expressed GFP, we could directly observe the influence of additional cell volume on the transcription of a single gene. We found that absolute GFP mRNA counts increased in fused cells as compared to the original small cells (Fig. 3.1), showing that increasing total cellular content is by itself sufficient to increase absolute mRNA abundance.

Moreover, the GFP mRNA counts scaled with heterokaryon volume (Fig. 3.2), although with a different concentration than in the unfused cells, suggesting that the rate of GFP transcription scaled with the ultimate volume of the fused cell. The fact that the nucleus from the WM983b-GFP-NLS cell could change its overall transcriptional activity showed that the modulation occured via the activity of a diffusible *trans*-acting factor. The alternative is that transcription levels are hard-coded into the DNA via some *cis*-acting factor. If this were the case, each nucleus would continue producing RNA at the same levels as before fusion, and we would not observe an increase in GFP mRNA upon fusion.

How might such a *trans* factor transmit volume information to the GFP gene in order to increase its transcription concordantly with the increase in cellular volume? There are two broad categories of mechanism: (1) The factor acts as a "volume sensor" and does not know about the amount of DNA in the cell. An example could be a modifiable global transcription factor protein whose degree of modification/activity is

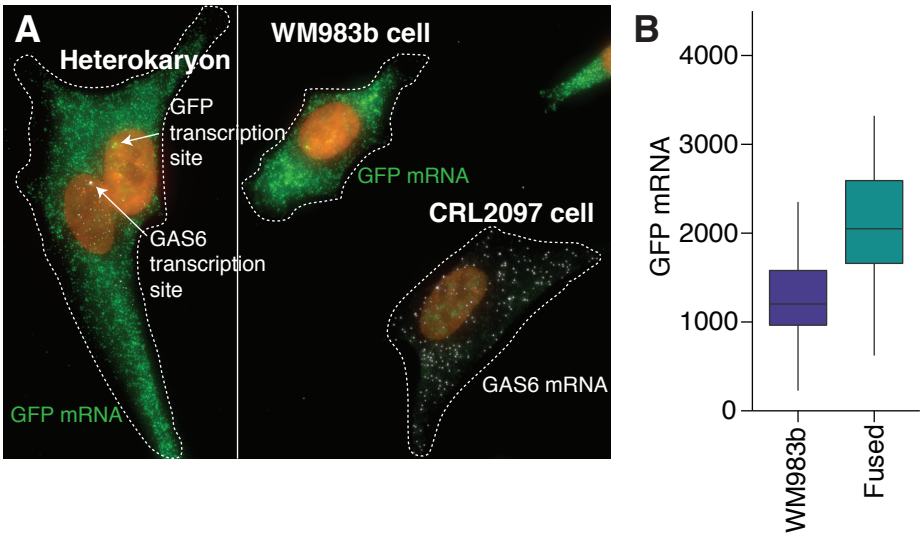

Figure 3.1: GFP mRNA is expressed at higher levels in fused cells. (A) Representative image of fused cells (heterokaryon, left) and unfused cells (WM983b, primary fibroblast, right). DAPI stain is in orange, GFP mRNA is in green, and *GAS6* mRNA is in white. White arrows indicate transcription sites. (B) Quantification of GFP mRNA in unfused and fused cells. Box extends to first and third quartile, and whiskers extend to the maximum-distance points within 1.5 inter-quartile ranges of the box. Data are a combination of two biological replicates.

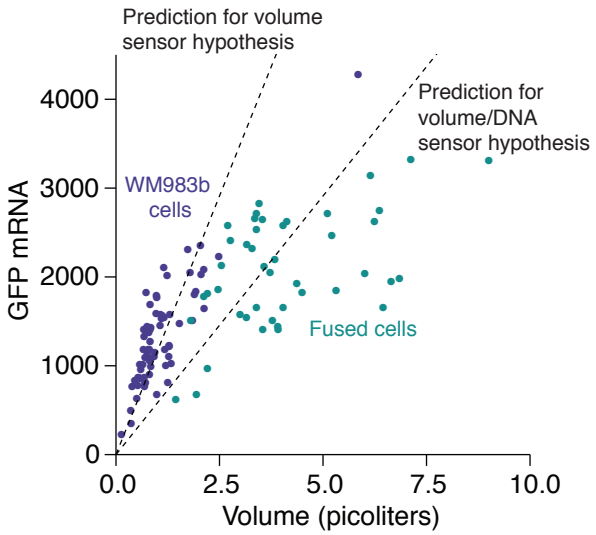

Figure 3.2: GFP mRNA scales with volume in fused cells. GFP vs. volume for fused and unfused cells. Upper dashed line represents best fit line by linear regression for unfused cells. Lower dashed line is a model, not fit, and has a slope that is half of the upper fit line.

proportional to cellular size (Fig. 3.3, left). The role of this factor is essentially to communicate to the nucleus the size of the cell and have it produce RNA accordingly. (2) The factor acts as a "volume/DNA sensor" whose activity depends on both cellular volume and DNA content. One such mechanism is the existence of a general transcription factor of limiting abundance relative to the number of binding sites in the DNA (limiting factor, Fig. 3.3, right). We assume that the factor is expressed proportional to cellular volume, so there is a higher absolute amount of factor in large cells than in small cells. Here, the DNA "counts" how big the cell is by binding all available factor molecules, thus increasing transcription in bigger cells as more factor binds to DNA. Another possibility is the sequestration of such a factor to the nucleus, which can also achieve the same volume/DNA sensing behavior if nuclear volume is only weakly dependent on cellular volume. In this second case, the factor need not be limiting relative DNA binding sites. We explain this model, and some interesting

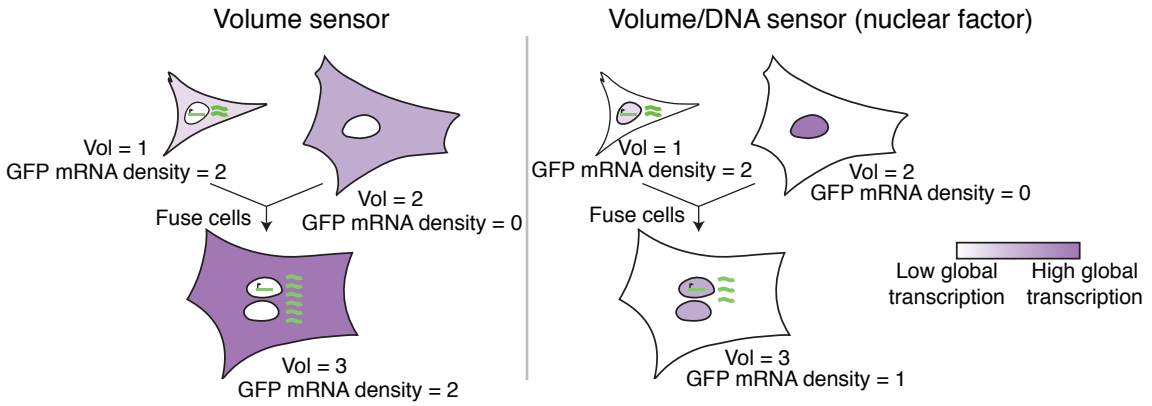

Figure 3.3: Models of transcriptional output in fused cells.

consequences thereof, in Section 3.3.

We distinguished between these two alternatives by comparing the concentration of GFP mRNA in the fused and unfused cells (Fig. 3.2). In the volume sensor scenario, the fusion cell would have the same concentration of GFP mRNA as the original small cells because the factor transmits the volume information to the GFP gene independent of the number of nuclei in the cell. We expect that transcription from each nucleus would be the same as it would be if there were only a single nucleus in the new, larger volume. While GFP mRNA would display the same concentration in this scenario because the GFP gene is present in only one nucleus, we might expect that genes present in both nuclei would have double the mRNA concentration, as both nuclei would produce this mRNA to fill the volume without knowledge of the other nucleus.

In the volume/DNA sensor scenario, the fusion cell would have half the concentration of GFP mRNA because the factor senses both the increased volume and the two nuclei. For example, a limiting factor would be diluted between the two nuclei in the fused cell, and each nucleus would only produce half the mRNA as it would

if it were the sole nucleus in a cell of that size. In this scenario, we expect GFP mRNA concentration to be half, but the concentration of mRNA from genes present in both nuclei (e.g. *GAPDH*) to be the same as in unfused cells. We found that the concentration of GFP mRNA in the fused cells was strictly less than and very close to half the concentration in unfused cells. We conclude that the factor responsible for increased transcription in smaller cells was not a volume sensor, but responded to both the size and DNA content. These results suggest that perturbations that change cell size will indirectly change global transcript counts per cell through this generic mechanism.

3.2 Transcriptional burst size increases in larger cells

We next sought to understand the mechanism by which the diffusible *trans* factor described above affected transcription by further by examining the relationship between volume and transcription of individual genes. mRNA is produced in bursts, marked by bright accumulations of nascent mRNA at the site of transcription. We characterize this bursting behavior through burst size (how much RNA is produced during a single burst) and burst fraction (how often a gene is actively transcribing, which is related to burst frequency [29]). Refer to Section 2.5 for a detailed explanation of our measurements. To quantify burst size, we measured the intensity of transcription sites for four genes in our fibroblast cells, and found higher intensity transcription sites in larger cells (Fig. 3.4), indicating that the factor we have described works by modulating transcriptional burst size. We note that transcriptional burst fraction was similar in cells of all volumes (Fig. 3.7), so the main source of transcriptional modulation across cells of different sizes results from a change in burst size. We further

showed that transcriptional burst size does not change throughout the cell cycle (Fig. 3.4), showing that there is a direct link between transcriptional burst size and cellular volume.

We hypothesize that this volume/DNA sensing *trans* factor, which may be either a limiting factor or one that is simply sequestered to the nucleus (described in detail in Section 3.3), has the following properties: (1) it is important for transcription, (2) is is almost purely nuclear, and (3) the total amount of the factor is proportional to the volume of the cell. The general transcriptional machinery, of which RNA polymerase II is a key component, satisfies these requirements. We confirmed that by reducing the amount of RNA polymerase II in the cell, we saw a reduction in transcription site intensity. This intensity is proportional to burst size [29], although we note that saturation may limit the dynamic range of this measurement [56]. We treated primary fibroblast cells with 100nM triptolide, which targets RNA polymerase II for degradation [3], reducing its levels (Fig. 3.5). After one hour, we saw a reduction of bright transcription sites for two different genes, showing that transcription site intensity depends directly on the amount of transcriptional machinery available. We note that in both knockdown and control conditions, there are many transcription sites with similar low intensities. We believe this intensity is representative of the production of a single transcript, which should remain the same between the two conditions. By reducing the amount of polymerase, we reduced the number of transcription sites producing more than one transcript per burst. We further note that transcription site fraction did not change significantly between the two conditions (Fig. 3.5), demonstrating that changing the amount of transcriptional machinery changes only burst size, not burst frequency.

The Churchman lab at Harvard performed a fractionation assay for us [37], showing that RNA polymerase II is 94% nuclear (Fig. 3.6), further supporting the hypothesis

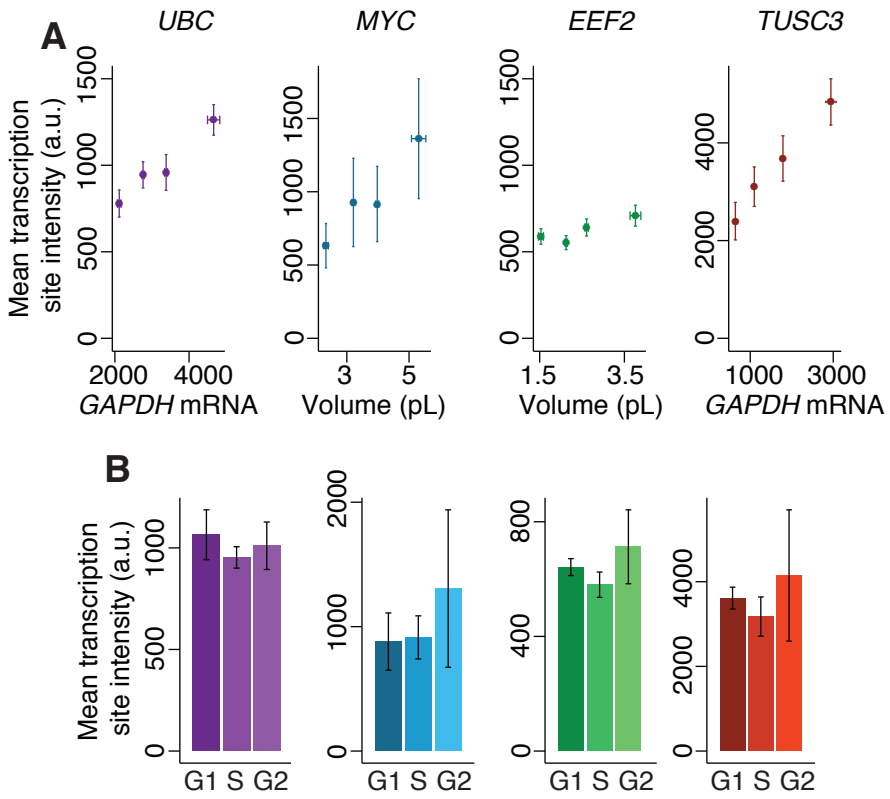

Figure 3.4: Transcription site intensity increases with volume, but not cell cycle stage. (A) Transcription site intensity and volume in primary fibroblast cells for genes *UBC*, *MYC*, *EEF2*, and *TUSC3*. Each data point represents the mean transcription site intensity per cell for a quartile of cells classified by volume or *GAPDH*. We detected transcription sites through intron/exon colocalization using RNA FISH. We calculated volume for *EEF2* data using *EEF2* as a guide, and volume for *MYC* data using *GAPDH*. We use *GAPDH* mRNA count as a proxy for volume for *UBC* and *TUSC3*. (B) Transcription site intensity and cell cycle stage in primary fibroblast cells. We determined cell cycle stage by Cyclin A2 and the histone 1H4E mRNA counts (see Appendix A and Fig. 2.5). For intensity measurements, data for *UBC*, *MYC*, and *EEF2* are from one of two biological replicates (*EEF2*: n = 190, *UBC*: n = 202, *MYC*: n = 103 transcription sites). Data for *TUSC3* are combined from two biological replicates (n = 255 transcription sites).

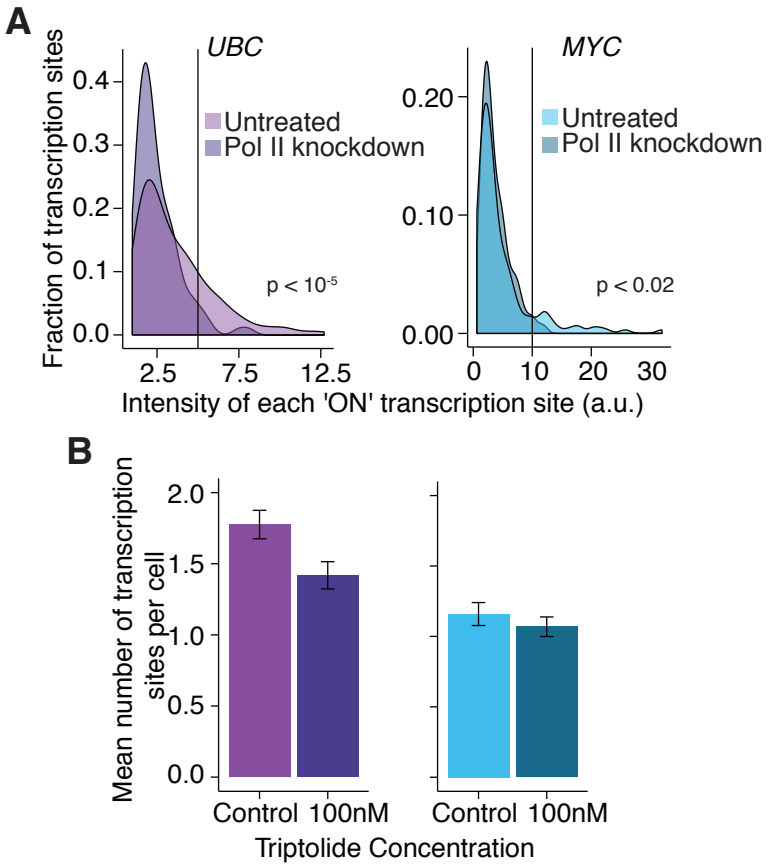

Figure 3.5: Burst size, but not fraction, decreases upon reduction of RNA polymerase. (A) Quantification of transcription site intensity before and after treatment with 100nM triptolide for one hour. P-value represents the probability of randomly finding the distributions of bright transcription sites (values to the right of the black line) in each condition. (B) Quantification of transcription site fraction (mean number of transcription sites per cell, not gated for cell cycle) before and after treatment with 100nM triptolide for one hour.

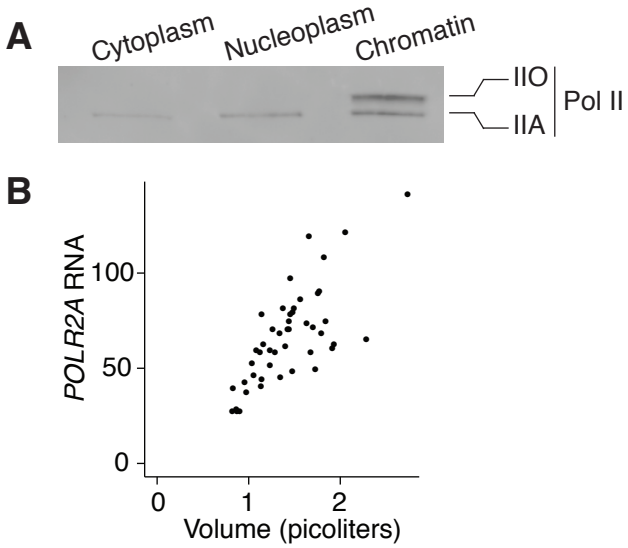

Figure 3.6: RNA polymerase is expressed proportional to volume and is almost entirely nuclear. (A) Western blot analysis reveals that >99% of the C-terminal domain hyper-phosphorylated form of RNA Polymerase II (IIO) is present in the chromatin fraction. The hypo-phosphorylated form of Pol II (IIA) is captured in all cellular fractions. We generated subcellular lysates from the same batch of primary fibroblast cells and probed with the F-12 antibody (Santa Cruz Biotechnology) that is directed against the N-terminal region of *RPB1*, the largest subunit of RNA polymerase II. We adjusted sample volumes so that Western blot signals of the subcellular fractions are comparable. (B) Quantification of RNA polymerase II mRNA (quantified by RNA FISH) vs. cytoplasmic volume in A549 cells. Data shown is from a single biological replicate.

that the general transcription machinery could be the diffusible *trans* factor we have been describing. Lastly, we performed RNA FISH on mRNA from the *POLR2A* gene, which encodes the large subunit of RNA polymerase II (Fig. 3.6). We saw that *POLR2A* mRNA scaled with volume, suggesting that polymerase is expressed proportional to cellular volume. Interestingly, unlike the mRNA from many housekeeping genes, the intercept term for *POLR2A* mRNA vs. volume is negative instead of positive, indicating that the concentration of polymerase is actually slightly higher in larger cells than it is in smaller cells. We do not fully understand why this is the case or how the cell might produce polymerase in such a manner, but we believe it may be a way for the cell to compensate for the reduction in concentration of nuclear proteins that comes from having slightly larger nuclei in larger cells (see next section).

3.3 Model of diffusible *trans* factor for volume/DNA ratio sensing

Here, we outline a fairly generic model for how a diffusible *trans* factor may transmit information on the ratio of volume to DNA to lead to increased transcription in larger cells irrespective of DNA content. The primary assumptions are that the factor is predominantly localized to the nucleus, the factor is required for mRNA transcription, and the cellular concentration of the factor is roughly constant irrespective of cellular volume (i.e., the total amount of factor is proportional to cellular volume). RNA polymerase II holoenzyme satisfies these conditions, although we do not claim that RNA polymerase II is the factor.

The model assumes binding of the factor to the DNA, and that only bound factor can result in productive transcription. The goal of the model is to provide a basis for the empirical finding that larger cells have increased transcription from the same absolute

number of DNA molecules. Our model encompasses two broad categories of mechanism that would lead to a perfectly linear scaling of transcription with cytoplasmic volume: (1) The factor is sequestered entirely in the nucleus, and so if the nucleus doesn't change with cellular volume, the concentration of the factor in the nucleus will be proportional to the total amount of factor. Thus, the factor will be proportionally more bound to the DNA in a larger cell than a smaller cell, producing more transcription. (2) The factor is a purely "limiting" factor in the sense that it has a very high affinity for DNA and the number of binding sites exceeds the amount of factor. In this situation, essentially all available factor will be bound to the DNA, and so for each gene, there would be proportionally more transcription in larger cells because more factor would be bound to DNA. These mechanisms are not necessarily mutually exclusive. The model incorporates affinity and nuclear volume as parameters, and so encompasses both of these potential mechanisms.

Briefly, the conclusion we derive from our model is that both scenarios pose viable mechanisms for scaling transcription with cellular volume. That said, we overall mildly favor scenario 1. Our data show that nuclear volume increases somewhat with nuclear size, which the model predicts should lead to a slight decrease in transcription in larger cells, and thus a higher concentration of mRNA in smaller cells, which is precisely what we observe. Moreover, there is a rough quantitative agreement between the degree of increased nuclear volume and the higher concentration of mRNA in smaller cells.

We begin with a few definitions. We use quantities within brackets to denote concentration (molecules per volume) and quantities without brackets to denote number of molecules per cell. For instance, $\text{factor}_{\text{free}}$ is the number of free molecules of the factor, while $[\text{factor}_{\text{free}}]$ is the number of free molecules per unit volume. $\text{factor}_{\text{DNA}}$ denotes the number of factor molecules instantaneously bound to DNA, $\text{factor}_{\text{total}}$ is the total amount of factor in the cell/nucleus, and DNA is the number of binding

sites on the DNA for the factor in the nucleus. K_{DNA} is the binding affinity of the factor for a particular gene. The cellular volume is given by V, and the nuclear volume by V_{nuclear}. Thus, given our assumption of proportionality, we define p_{factor} to be a constant such that $\mathrm{factor}_{\mathrm{total}} = p_{\mathrm{factor}} V$.

The total factor is given by

$$\mathrm{factor}_{\mathrm{total}} = \mathrm{factor}_{\mathrm{free}} + \mathrm{factor}_{\mathrm{DNA}}. \tag{3.1}$$

Dividing by the nuclear volume, we arrive at a relationship between concentrations:

$$[\mathrm{factor}_{\mathrm{total}}] = [\mathrm{factor}_{\mathrm{free}}] + [\mathrm{factor}_{\mathrm{DNA}}]\,. \tag{3.2}$$

The binding affinity is defined via mass action as

$$K_{\mathrm{DNA}} = \frac{[\mathrm{factor}_{\mathrm{free}}]\,[\mathrm{DNA}]}{[\mathrm{factor}_{\mathrm{DNA}}]}, \tag{3.3}$$

and may be different for different genes owing to promoters having different numbers of binding sites for the factor or different binding affinities.

Thus, the total concentration of the machinery bound to DNA is

$$[\mathrm{factor}_{\mathrm{DNA}}] = \frac{([\mathrm{factor}_{\mathrm{DNA}}] - [\mathrm{factor}_{\mathrm{total}}])\,[\mathrm{DNA}]}{K_{\mathrm{DNA}}}. \tag{3.4}$$

Solving for $[\mathrm{factor}_{\mathrm{DNA}}]$, we find:

$$[\mathrm{factor}_{\mathrm{DNA}}] = \frac{[\mathrm{factor}_{\mathrm{total}}]\,[\mathrm{DNA}]}{K_{\mathrm{DNA}} + [\mathrm{DNA}]}. \tag{3.5}$$

In the limiting case where $K_{\mathrm{DNA}} = 0$, we expect all of the factor to be bound to DNA, and in that case, we find $[\mathrm{factor}_{\mathrm{DNA}}] = [\mathrm{factor}_{\mathrm{total}}]$, as expected.

Relating concentrations to volumes yields

$$[\text{factor}_{\text{total}}] = \frac{p_{\text{factor}} V}{V_{\text{nucleus}}}, \tag{3.6}$$

where p_{factor} is the proportionality constant defined earlier. Similarly,

$$[\text{DNA}] = \frac{\text{DNA}}{V_{\text{nucleus}}}. \tag{3.7}$$

Hence,

$$[\text{factor}_{\text{DNA}}] = \frac{p_{\text{factor}} \frac{V}{V_{\text{nucleus}}} \frac{\text{DNA}}{V_{\text{nucleus}}}}{K_{\text{DNA}} + \frac{\text{DNA}}{V_{\text{nucleus}}}}. \tag{3.8}$$

Simplifying,

$$[\text{factor}_{\text{DNA}}] = \left(\frac{1}{V_{\text{nucleus}}}\right) \frac{p_{\text{factor}} \cdot V \cdot \text{DNA}}{K_{\text{DNA}} \cdot V_{\text{nucleus}} + \text{DNA}}. \tag{3.9}$$

Because $[\text{factor}_{\text{DNA}}] = \text{factor}_{\text{DNA}}/V_{\text{nucleus}}$, we can solve for the total amount of transcriptional machinery bound to DNA:

$$\text{factor}_{\text{DNA}} = \frac{p_{\text{factor}} \cdot V \cdot \text{DNA}}{K_{\text{DNA}} \cdot V_{\text{nucleus}} + \text{DNA}}. \tag{3.10}$$

In the limiting case $K_{\text{DNA}} = 0$ here, we find that $\text{factor}_{\text{DNA}}$ is directly proportional to volume, and equal to the total amount of factor in the nucleus irrespective of nuclear volume. However, in the case where K_{DNA} is not zero, then the volume of the nucleus will result in deviations from perfect scaling of transcription with cellular volume. Intuitively, if the volume of the nucleus increases somewhat in larger cells, then the concentration of the factor and the DNA will decrease and hence the amount of factor bound to DNA will be somewhat less than it would be otherwise. In that

case, larger cells would have somewhat less transcription than would be expected in the case of perfect scaling of transcription with cellular volume, which fits with our experimentally observed volume-independent transcript abundance (i.e., decreased mRNA concentration in larger cells). We also observed that nuclear volume is somewhat greater in larger cells. Thus, it was possible, at least qualitatively, that the increase in nuclear volume could explain the apparent decrease in mRNA concentration in larger cells. We thus wanted to check whether there is a quantitative agreement between our observed relationship between nuclear volume and cytoplasmic volume and the increased mRNA concentration in smaller cells, which would establish the plausibility of such a model.

As mentioned, our measurements show that nuclear area and cellular volume positively correlate. Approximating nuclear volume by raising nuclear area to the 3/2 power, we find a linear relationship between nuclear "volume" and cellular volume ($V_{\text{nucleus}} \propto a + bV$), with y-intercept $a = 2169$ fL (95% C.I. = (1923, 2381)fL), and slope $b = 0.9354$ (95% C.I. = (0.8313, 1.042)). It is important to note that the while the relationship is well fit by a line, the line does not pass through zero, and so nuclear volume is *not* directly proportional to total cellular volume. Using this linear relationship, we can express the total amount of factor bound to DNA as a function of cellular volume:

$$\text{factor}_{\text{DNA}}(V) = \frac{p_{\text{factor}} \cdot V}{\left(\tilde{a} + \tilde{b}V\right) + 1}, \tag{3.11}$$

where $\tilde{a} = \frac{K_{\text{DNA}}}{\text{DNA}} \cdot a$ and $\tilde{b} = \frac{K_{\text{DNA}}}{\text{DNA}} \cdot b$. The ratio a/b $(= \tilde{a}/\tilde{b})$ has units of volume, and is geometrically equivalent to the x-intercept of the line of best fit between nuclear volume and cellular volume.

We now wanted to check whether the volume-independent transcription we observed in our mRNA-volume plots would quantitatively agree with this model. Because the

factor is required for transcription and only transcribes when bound to DNA, then each gene essentially grabs a fixed proportion of the amount of factor bound to DNA. (This fraction will depend on the specific regulation of the gene.) So the total transcription of a gene will be proportional to $\text{factor}_{\text{DNA}}$. Thus, lumping together this proportionality constant along with mRNA production and degradation and other associated constants into a constant c, the relationship between RNA and volume is given by:

$$\text{RNA}(V) = c \cdot \text{factor}_{\text{DNA}}(V). \tag{3.12}$$

We should be able to fit our RNA vs. volume data using the above equation to obtain estimates for $\tilde{a}$ and $\tilde{b}$, in particular their ratio, which is directly comparable to the ratio a/b.

We did so for three genes and found fitting parameters:

	UBC	$ZNF444$	$EEF2$
$\tilde{a}$ (fL)	1.578	95.68	0.2747
95% C.I. ($\tilde{a}$) (fL)	(0.8524, 2.461)	(70.23, 127.9)	(-0.1518, 0.5630)
$\tilde{b}$	1.936×10^{-4}	0.01495	0.0003658
95% C.I. ($\tilde{b}$)	(-7.617×10^{-5}, 4.595×10^{-4})	(0.004163, 0.02394)	(0.0002680, 0.0005879)
$\tilde{a}/\tilde{b}$ (fL)	7044	6446	744.4
95% C.I. ($\tilde{a}/\tilde{b}$) (fL)	(-62743, 111200)	(2988, 28110)	(-253.5, 2064)

For the fit of nuclear area to volume, we find the ratio $a/b = 2329$ fL, with a 95% confidence interval of (1844, 2851) fL. We note that the ratios $\tilde{a}/\tilde{b}$ for all of our genes are of that same order of magnitude, albeit with large error. This result suggests that the above equation for $\text{factor}_{\text{DNA}}(V)$ may be the equation governing the production of

mRNA in cells. This model provides an explanation that is quantitatively consistent with our data for why smaller cells exhibit proportionally slightly more transcription than larger cells—nuclei in small cells are slightly smaller than those in large cells, increasing the concentration of $\text{factor}_{\text{DNA}}(V)$, and therefore increasing transcription.

Our results are consistent with RNA polymerase II holoenzyme being the factor. RNA polymerase II is required for transcription, transcribes when bound to DNA, and is almost exclusively localized to the nucleus. Also, most reports indicate that most RNA polymerase II in the nucleus is not specifically bound to DNA [9, 26, 27]. Based on that fact, one would expect that increased nuclear size should lead to slight under-transcription, as we observed. Our analysis shows that this relationship is quantitatively plausible. Further studies will be required to rigorously establish that RNA polymerase II holoenzyme is the factor that connects volume/DNA ratio to transcription.

We have shown that the RNA encoding the large subunit of RNA polymerase II is expressed proportionally to volume, but with a negative intercept term, suggesting that the concentration of RNA polymerase may be higher in larger cells. We do not understand mechanistically how this might happen, but we suspect it may be a way for larger cells to compensate for having larger nuclei. By producing more polymerase in larger cells, larger cells will have a slight boost in transcription. Our observations show that large cells do have a slightly lower concentration of RNA than smaller cells, so the higher levels of polymerase are not a complete compensation, but still may be a means of keeping concentrations more similar between large and small cells.

Note, however, that in many situations such as in early embryogenesis, nuclear size does change dramatically as a function of cellular volume. In these situations, the mechanism we describe would face a challenge because the concentration of RNA polymerase II holoenzyme in the cell's nucleus would remain the same after division

(and the associated decrease in cellular volume), leading to over-transcription. The limiting factor model (scenario 2, with K_{DNA} very small) would not suffer from these issues. It is possible that some intermediate scenario is at play in early embryogenesis.

Another problem with these models is the potential for runaway positive feedback, in which a random increase in the factor would lead to more production of the factor, thus leading to runaway transcription. For this reason, we expect that the cell maintains strong control of factor levels to avoid these issues. Ultimately, a complete understanding of factor dynamics will likely require adding growth to models of transcriptional homeostasis.

3.4 A DNA-linked *cis*-acting factor reduces transcription fraction, not burst size, immediately after DNA replication

Thus far, we have discussed the means by which large cells and small cells produce different amounts of mRNA from the same amount of DNA. However, cells have a second challenge to overcome as well. We have shown that cells in the G1 and G2 phases of the cell cycle can have the same volume and same mRNA concentration (Fig. 2.6), although cells in G2 have twice the number of DNA molecules as those in G1. How then do two cells of the same size produce the same amount of mRNA, despite having different amounts of DNA?

We previously found that transcriptional burst size scales with volume (Fig. 3.4), so we now measured burst fraction. To measure transcriptional burst fraction, we counted active transcription sites in each cell and divided by the total number of gene copies for each stage of the cell cycle (two copies in G1, four copies in G2). For

all of the genes we measured, the fraction per gene copy in G2 was approximately half of that in G1 (Fig. 3.7). This is a very interesting result, as it shows that cells can produce the same amount of mRNA before and after DNA replication by only allowing each gene copy to transcribe with half the frequency after replication. We note that this is not due to repression of replicated copies of DNA, as we observed cells in G2 with four active transcription sites (Fig. 3.8). This showed that the cell has a mechanism to precisely reduce transcription frequency in G2 to keep overall transcription constant across the G1 and G2 phases of the cell cycle. Transcription burst fraction did not change with volume (Fig. 3.7), and transcription burst size did not change over the course of the cell cycle (Fig. 3.4), showing that this mechanism is distinct from the volume-compensating mechanism described above.

We were surprised that the mechanism compensating for DNA differences over the cell cycle was different from the one compensating for volume. In principle, limiting factor models, as described above, that may be responsible for scaling transcription with volume would also compensate for increased gene copy number due to DNA replication. If the limiting factor was completely bound to DNA before and after DNA replication, only half as much factor would be bound to each gene copy in G2 as in G1. This could theoretically lead to a decrease in transcription frequency by half in G2, although it would also predict a decrease in transcriptional burst size in G2, which we do not observe (Fig. 3.7). Moreover, such models predict an inappropriate boost in transcription for genes that replicate early in S phase. The limiting factor would distribute itself over all the DNA, essentially "double counting" the small percentage of genes that replicate considerably earlier than the majority of DNA (Fig. 3.9). If this were the case, we would expect early replicating genes to have similar transcription frequencies per gene copy in G1 and S phase, although each gene would be present in four copies in S phase instead of two as in G1, essentially doubling the transcriptional

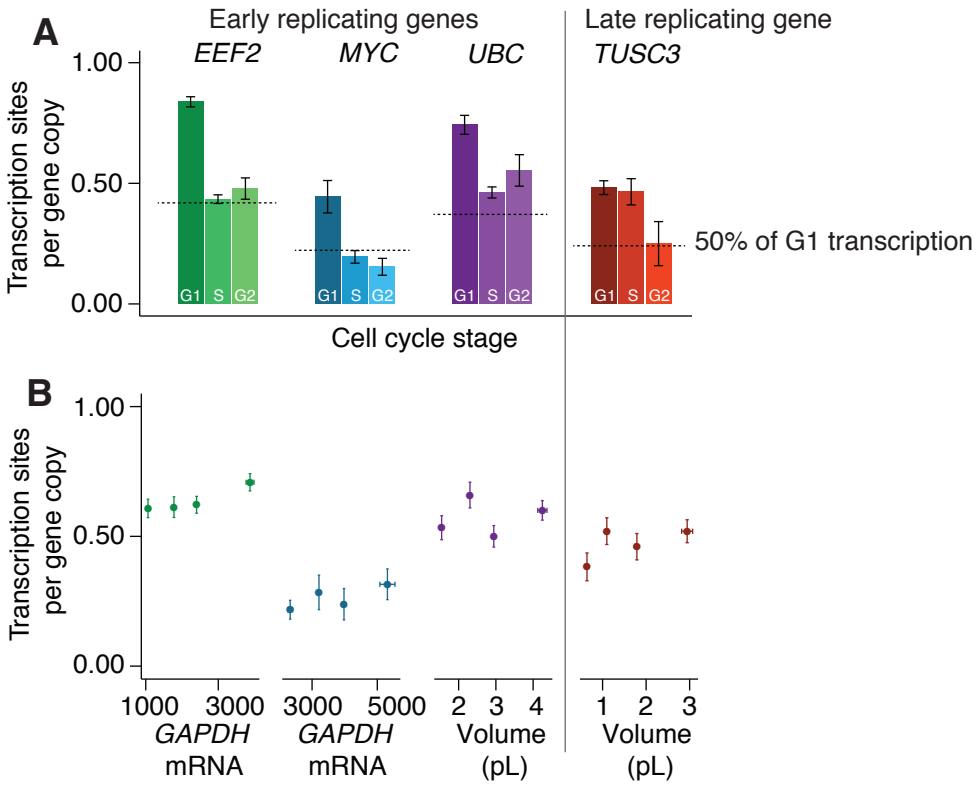

Figure 3.7: A *cis*-acting factor decreases transcription frequency immediately upon DNA replication. (A) Fraction of transcription sites by cell cycle stage in primary fibroblast cells. We determined cell cycle stage by Cyclin A2 and Histone 1, H4E mRNA counts. Dashed lines represent half the fraction of active transcription sites in G1. We normalize all G1 data to two gene copies, and all G2 data to four gene copies. For *EEF2*, *MYC*, and *UBC* (early replicators), we normalize S phase data to four gene copies. For *TUSC3* (late replicator), we normalize S phase data to two gene copies. (B) Number of transcription sites per gene copy classified by volume in primary fibroblast cells. Each data point represents the mean number of transcription sites for a quartile of cells classified by volume. We calculated volume for *EEF2* data using *EEF2* as a guide, and volume for *MYC* data using *GAPDH*. We use *GAPDH* as a proxy for volume for *UBC* and *TUSC3*. For burst fraction measurements, data for *EEF2*, *UBC*, and *TUSC3* are a combination of two biological replicates (*EEF2*: n = 516, *UBC*: n = 332, *TUSC3*: n = 255 transcription sites). Data for *MYC* is from one of two biological replicates (*MYC*: n = 103 transcription sites).

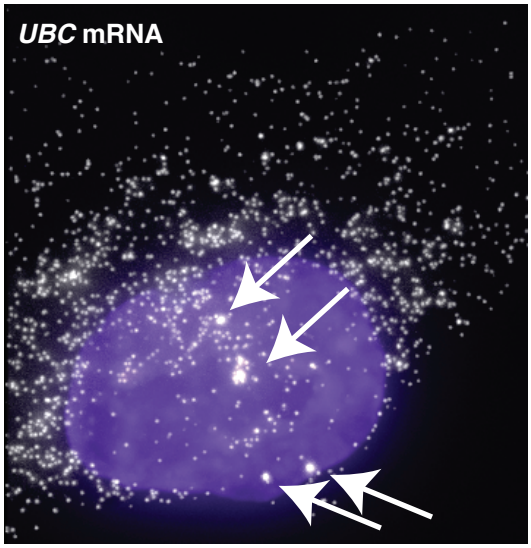

Figure 3.8: Replicated gene copies are transcriptionally competent. *UBC* mRNA in a primary human fibroblast cell. RNA FISH probe in white, DAPI stain in purple. White arrows indicate transcription sites. We detect transcription sites through intron/exon colocalization by RNA FISH. This cell is in G2 and has four transcription sites, demonstrating that all gene copies are transcriptionally competent after replication.

output between G1 and S phase.

To see if this was the case, we measured transcriptional burst fraction for early replicating genes in S phase (*EEF2*, *MYC*, *UBC*; see Fig. 3.10 for replication timing). We classified cells as being in G1, S, or G2 based on cell cycle markers as described in Fig. 2.5. These early replicating genes all showed the same transcription fraction per gene copy in S and G2, implying that overall transcriptional output of these genes remained the same throughout the cell cycle. This observation ruled out the possibility that transcription fraction changes simply because a factor gets diluted between copies of replicated DNA.

This leaves two alternatives for burst frequency reduction between G1 and G2: (1) transcription frequency could universally decrease by a factor of two upon the initiation of S phase, or (2) transcription frequency could decrease on a gene-by-gene

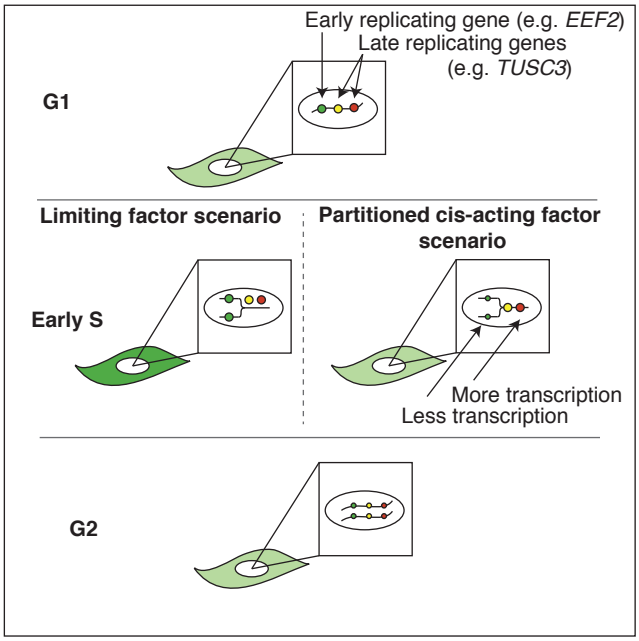

Figure 3.9: Schematic of potential mechanisms for changing gene expression with cell cycle.

basis (i.e., in *cis*) immediately upon DNA replication. We note that scenario 1 would actually lead to the opposite problem of the one described above—genes that replicate late in S phase would be *under*-transcribed for the majority of S phase. Because such genes only have two copies for the majority of S phase, if frequency is reduced at the beginning of S phase, these genes would have half the transcriptional output in S as in G1.

To test between these alternatives, we imaged transcription of a gene that replicates very late in S phase (*TUSC3*; Fig. 3.10). If transcription frequency were universally reduced by a factor of two at the beginning of S phase, this gene would have the same transcription fraction per gene copy in S and G2, despite having two gene copies in S and four in G2. However, we found that this late-replicating gene maintains G1 levels of transcription through S phase, and does not reduce transcription fraction until G2. Therefore we conclude that there exists a mechanism whereby transcription frequency

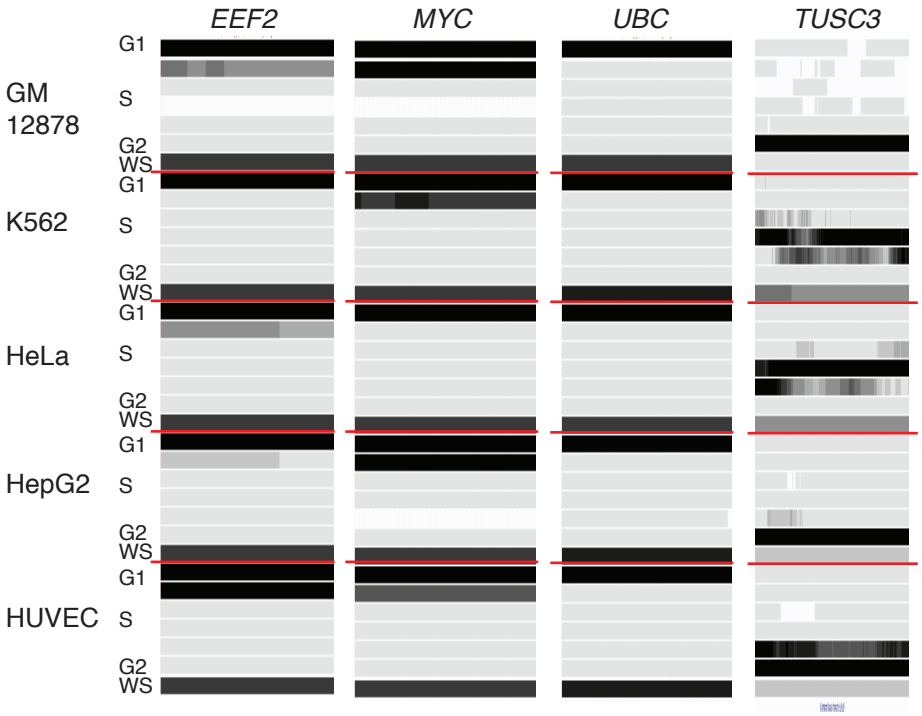

Figure 3.10: *EEF2*, *MYC*, and *UBC* genes are replicated early in the cell cycle; *TUSC3* replicates late. Tracks from UCSC genome browser displaying UW Repli-Seq data in GM12878 (lymphoblastoid), K562 (chronic myelogenous leukemia), HeLa (cervical cancer), HepG2 (liver carcinoma), and HUVEC (human umbilical vein endothelial) cell lines. The track displays data for different points in the cell cycle: G1, S1 (early S phase), S2 (middle-early S phase), S3 (middle-late S phase), S4 (late S phase), and G2. WS represents a wavelet-smoothed transform of the six other tracks. This data was generated by sequencing newly-replicated DNA in each point in the cell cycle. Darkness of track corresponds to read density.

is reduced by a factor of two immediately upon replication of that gene, with different timing for different genes. Candidates for such a mechanism include the partitioning or modification of DNA-linked factors, such as histones with particular modifications, upon DNA replication, resulting in half the transcriptional burst frequency as before replication.

An interesting question is how this mechanism "resets" after a cell divides. After cell division, each new daughter cell gets two of the four copies of DNA from the mother cell. If nothing changed during division, each of the daughter cells would still transcribe with the G2 frequencies, leading to an under-production of mRNA. To overcome this, each cell must return to its original G1 transcription frequency between M phase and the new G1 phase after division. It is likely that histones and basal histone marks stay in place during M phase, although there is evidence for histone acetylation and methylation occurring during M phase [68]. It could be that histone modifications are added or removed during M phase to reset transcription frequency back to an appropriate level. How transcription frequency is reset is an interesting question, and one that remains open.

Together our data demonstrate the existence of two separate transcriptional mechanisms that allow cells to maintain RNA concentration despite changes in DNA content and cellular volume. Cells modulate transcriptional burst size through a *trans* mechanism to allow larger cells to produce more mRNA from the same amount of DNA, and modulate burst frequency over the cell cycle in *cis* to maintain RNA concentration despite changes in DNA content.

Chapter 4

But what about the rest of us? Exploring noise in RNA expression

Our RNA FISH data revealed that while the expression of most genes was consistent with a volume-dependent transcription rate, many genes showed strong variability in transcript concentration from cell to cell (see Appendix B, note *MYC*, *ICAM1*, *ACTA2*). There is a precedent for such variability in the literature, and indeed there is an entire scientific field dedicated to the study of variability in gene expression [31, 49, 51, 54].

Variability or "noise" in gene expression is typically defined by measuring levels of individual RNAs or proteins in single cells, without taking extrinsic factors such as volume into account. The typical noise measure is the coefficient of variation, $CV = \sigma/\mu$, which is simply the standard deviation of the levels of RNA or protein in single cells from a population normalized to the mean value across the population. Here we have shown that the size of a cell can greatly impact RNA production, leading to six-fold or greater differences in RNA levels between single cells. The coefficient of variation does not distinguish between variability from volume and variability from other sources. We were interested in quantifying noise from sources

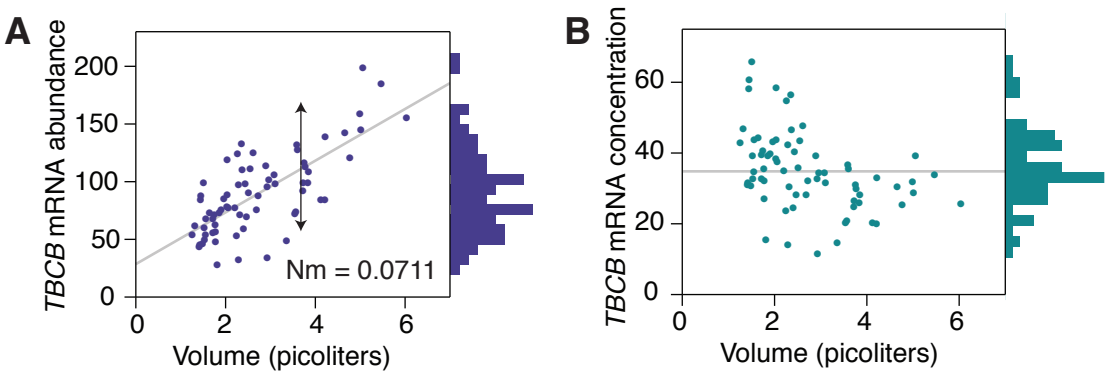

Figure 4.1: Visual representation of volume-corrected noise measure. (A) *TBCB* mRNA abundance and volume in primary fibroblast cells. Each point represents a single-cell measurement by RNA FISH. Histogram indicates mRNA distribution. Arrow indicates volume-corrected noise measure. Gray line is best linear fit. (B) *TBCB* mRNA concentration vs. volume. These data are the same as in (A), but each is normalized by volume. Histogram indicates distribution of mRNA concentration. Gray line indicates average concentration. Data are from a combination of two biological replicates.

other than volume. Our collaborator Abhyudai Singh, Assistant Professor of Electrical and Computer Engineering at the University of Delaware, developed a metric to quantify this variability, which we term the "volume-corrected noise measure" (Nm) [5, 25, 60], which is in principle similar to the squared coefficient of variation of mRNA concentration, but accounts for volume-independent transcription (Fig. 4.1). The following section (Section 4.1) was written in its entirety by Abhyudai Singh.

4.1 Computing volume-corrected noise measure from single-cell mRNA and volume measurements

We define volume-corrected noise measure as the cell-to-cell expression variability in mRNA levels that cannot be accounted for by cell-to-cell differences in volume.

Throughout the section, we denote the variance of a random variable X by σ_X^2.

Let m and V be random variables denoting single-cell mRNA level and volume, respectively. The expected number of mRNA transcripts in a cell given its volume V is assumed to increase linearly with V, i.e.,

$$\langle m|V\rangle = a + bV \implies \langle m\rangle = a + b\langle V\rangle, \tag{4.1}$$

where $\langle . \rangle$ represents the expected value, and a, b are gene-specific constants (related to volume-independent and volume-correlated transcript abundance, respectively). From (4.1), the covariance between m and V is given by

$$Cov(m, V) = \langle mV\rangle - \langle m\rangle\langle V\rangle = \langle (a + bV)V\rangle - (a + b\langle V\rangle)\langle V\rangle = b\sigma_V^2. \tag{4.2}$$

The extent of cell-to-cell variability in mRNA counts that can be accounted for by volume is

$$\sigma^2_{\langle m|V\rangle} = \sigma^2_{a+bV} = b^2\sigma_V^2, \tag{4.3}$$

which using (4.2) can be written

$$\sigma^2_{\langle m|V\rangle} = bCov(m, V). \tag{4.4}$$

Volume-corrected noise measure Nm is defined as

$$Nm := \frac{\sigma_m^2 - \sigma^2_{\langle m|V\rangle}}{\langle m\rangle^2} \tag{4.5}$$

is obtained as follows using (4.4)

$$Nm = CV_m^2 - \frac{bCov(m,V)}{\langle m\rangle^2} = CV_m^2 - S\frac{Cov(m,V)}{\langle m\rangle\langle V\rangle}, \tag{4.6}$$

where CV_m^2 represents the total variability in mRNA levels measured by its Coefficient of Variation (CV) squared and

$$S = \frac{b\langle V\rangle}{\langle m\rangle} = \frac{b\langle V\rangle}{a + b\langle V\rangle}. \tag{4.7}$$

4.1.1 Noise measure in a two-state promoter model

Consider two alleles, where each allele transitions independently between active and inactive states with rates k_{on} and k_{off}. We assume that the transcription rate from the active state increases linearly with cell volume V. We first compute CV_m^2 (mRNA coefficient of variation squared) for the case where transcription is independent of volume, and then extend it to the volume dependent case.

Transcription rate independent of volume

Let the transcription rate from active state be k_m. Then, the steady-state first and second-order moment of the mRNA level m is given by

$$\langle m\rangle = \frac{2G_{on}k_m}{\gamma_m}, \quad \langle m^2\rangle = \langle m\rangle + \frac{\gamma_m(1-G_{on})\langle m\rangle^2}{2(G_{on}\gamma_m + k_{on})} + \langle m\rangle^2, \tag{4.8}$$

where

$$G_{on} = \frac{k_{on}}{k_{on} + k_{off}} \tag{4.9}$$

is the fraction of time an allele is in the active state, and γ_m is the mRNA degradation rate. Note that the factor of two in (4.8) arises due to the presence of two alleles. This results in

$$CV_m^2 := \frac{\langle m^2\rangle - \langle m\rangle^2}{\langle m\rangle^2} = \frac{1}{\langle m\rangle} + \frac{\gamma_m(1-G_{on})}{2(G_{on}\gamma_m + k_{on})}. \tag{4.10}$$

Transcription rate dependent on volume

We assume $k_m = a + bV$, where volume V is a random variable with mean $\langle V\rangle$ and variance σ_V^2. Based on (4.8),

$$\langle m|V\rangle = \frac{2G_{on}(a+bV)}{\gamma_m}, \quad \langle m^2|V\rangle = \langle m|V\rangle + \frac{\gamma_m(1-G_{on})\langle m|V\rangle^2}{2(G_{on}\gamma_m + k_{on})} + \langle m|V\rangle^2. \tag{4.11}$$

Unconditioning on the volume we obtain

$$\langle m\rangle = \frac{2G_{on}(a+b\langle V\rangle)}{\gamma_m} \tag{4.12a}$$

$$\langle m^2\rangle = \langle m\rangle + \frac{\gamma_m(1-G_{on})\left\langle\langle m|V\rangle^2\right\rangle}{2(G_{on}\gamma_m + k_{on})} + \left\langle\langle m|V\rangle^2\right\rangle. \tag{4.12b}$$

Using (4.11)

$$\left\langle\langle m|V\rangle^2\right\rangle = \left\langle\left(\frac{2G_{on}(a+bV)}{\gamma_m}\right)^2\right\rangle = \langle m\rangle^2(1+S^2CV_V^2), \tag{4.13}$$

where the mean mRNA count $\langle m\rangle$ is given by (4.12a), S is given by (4.7) and CV_V^2 is the volume CV^2. Substituting (4.13) in (4.12b)

$$\langle m^2\rangle = \frac{\gamma_m(1-G_{on})\langle m\rangle^2(1+S^2CV_V^2)}{2(G_{on}\gamma_m + k_{on})} + \langle m\rangle^2(1+S^2CV_V^2) + \langle m\rangle. \tag{4.14}$$

Above equation yields

$$CV_m^2 := \frac{\langle m^2 \rangle - \langle m \rangle^2}{\langle m \rangle^2} = \frac{1}{\langle m \rangle} + \frac{\gamma_m(1 - G_{on})(1 + S^2 CV_V^2)}{2(G_{on}\gamma_m + k_{on})} + S^2 CV_V^2. \tag{4.15}$$

As expected, (4.15) reduces to (4.10) when $CV_V^2 = 0$. In (4.15), the first term represents Poissonian noise in mRNA levels due to random birth and death of individual mRNA molecules. The second term is the noise contribution from stochastic promoter switching. Variation in mRNA levels due to cell-to-cell differences in cell volume is represented by the last term. Removing the last term, we obtain the noise measure as

$$Nm = \frac{1}{\langle m \rangle} + \frac{\gamma_m(1 - G_{on})(1 + S^2 CV_V^2)}{2(G_{on}\gamma_m + k_{on})}. \tag{4.16}$$

4.1.2 Estimating promoter transition rates between active and inactive states

Noise measures obtained from single-cell mRNA count and volume measurements are used to estimate promoter transition rates k_{on} and k_{off} using (4.16). To correct for measurement noise, we take into account a 15% error in mRNA counting. From (4.9) and (4.16)

$$G_{on} = \frac{k_{on}}{k_{on} + k_{off}} \tag{4.17a}$$

$$Nm = \frac{1}{\langle m \rangle} + \frac{\gamma_m(1 - G_{on})(1 + S^2 CV_V^2)}{2(G_{on}\gamma_m + k_{on})} + CV_{count}^2, \tag{4.17b}$$

where $CV_{count}^2 = 0.15^2 = 0.0225$ represents the mRNA counting error. In the above equations, quantities Nm, S (defined in (4.7)), CV_V^2 (volume CV^2), $\langle m \rangle$, G_{on} are computed from data for a given gene. Using mRNA half-life information from literature, rates k_{on} and k_{off} can be estimated by solving (4.17). The average promoter dwell

time in the active and inactive state is given by

$$T_{on} = \frac{1}{k_{off}}, \quad T_{off} = \frac{1}{k_{on}}, \tag{4.18}$$

respectively, and reported in Table I under the column "T_{on}, T_{off} from Nm". Since there may be other unaccounted sources of noise in gene expression, these dwell time estimates should be considered an upper bound on their actual values.

We contrast the above estimates to the scenario where all the mRNA expression variability is assumed to arises from transcriptional bursting. From (4.10), k_{on} and k_{off} in that case would be estimated by solving

$$G_{on} = \frac{k_{on}}{k_{on} + k_{off}} \tag{4.19a}$$

$$CV_m^2 = \frac{1}{\langle m \rangle} + \frac{\gamma_m(1 - G_{on})}{2(G_{on}\gamma_m + k_{on})} + CV_{count}^2. \tag{4.19b}$$

Since $CV_m^2 > Nm$, dwell times obtained from (4.19) (see "T_{on}, T_{off} from CV_m^2" in Table I) are significantly larger than those obtained from (4.17). For example, using the *GAPDH* noise measure, we estimate $T_{on} = 4.9$ hours. However, if one ignores the contribution of cell volume in driving intercellular variation in *GAPDH* mRNA, the mean dwell time in the active state is obtained to be $13 - 14$ days (331 hours) from (4.19).

4.2 Noise in RNA FISH measurements

We evaluated Nm for all of the genes we measured by RNA FISH. A standard measure of gene expression noise is the coefficient of variation (CV, standard deviation divided by mean), which only takes into account the spread and the mean of the expression data. Using such a measurement, most of the genes we measured by RNA

Average promoter dwell-time (4.18) obtained from the noise measure (Eq. (4.17)) or the total mRNA expression variability (Eq. (4.19)). mRNA half-lives and dwell times are reported in hours.

Gene	G_{on}	CV_m^2	Nm/CV_m^2	mRNA half-life	T_{on}, T_{off} from Nm	T_{on}, T_{off} from CV_m^2
ACTN4	0.65	0.14	0.4	13	6.1, 3.3	40.5, 21.8
GAPDH	0.65	0.23	0.17	24	4.9, 2.6	331, 178
EEF2	0.75	0.18	0.17	16	3.2, 1.1	1443.6, 481.2
FTL	0.3	0.58	0.32	24	6.3, 14.6	45.4, 105.9
ICAM1	0.05	0.8	0.8	6.5	0.5, 9.7	0.8, 15.4
ACTA2	0.2	1.18	0.9	3	5.1, 20.3	7.9, 31.6
LUM	0.15	0.7	0.75	24	7.8, 44.3	12.7, 72.5
SUPTH5	0.3	0.12	0.58	10	0.45, 1.1	1.4, 3.3

FISH would be deemed "noisy", or far from Poisson noise levels. We calculated CV^2 for mRNA counts as well as mRNA concentration (counts/volume) for many genes in our fibroblast cell line. Both of these measures assign higher noise levels to genes than the volume-corrected noise measure (Fig. 4.2). A Poisson process is the "least noisy" random process, in which the time between events (here, the time between production events of a single mRNA molecule) is exponentially distributed, and all events are independent. Because mRNA production occurs in bursts, we expect noise levels to be higher than Poisson, particularly if we are simply counting molecules and not accounting for volume. If, however, we instead measure Nm for each of our RNA FISH genes, we see that many of them display levels of variability near to or indistinguishable from Poisson (Fig. 4.3) once we account for measurement error (upper bound of around 15% [47]). Our measurement error represents how accurately we can detect and count single mRNAs. We plotted Nm against both mean mRNA abundance and mean mRNA half-life (Fig. 4.3, half-life values from [62]). Interestingly, Nm does not appear to depend significantly on either abundance or half-life, suggesting that there is no simple way to estimate noise measures of genes without actually measuring Nm.

Using RNA FISH, we were able to calculate noise measures for approximately

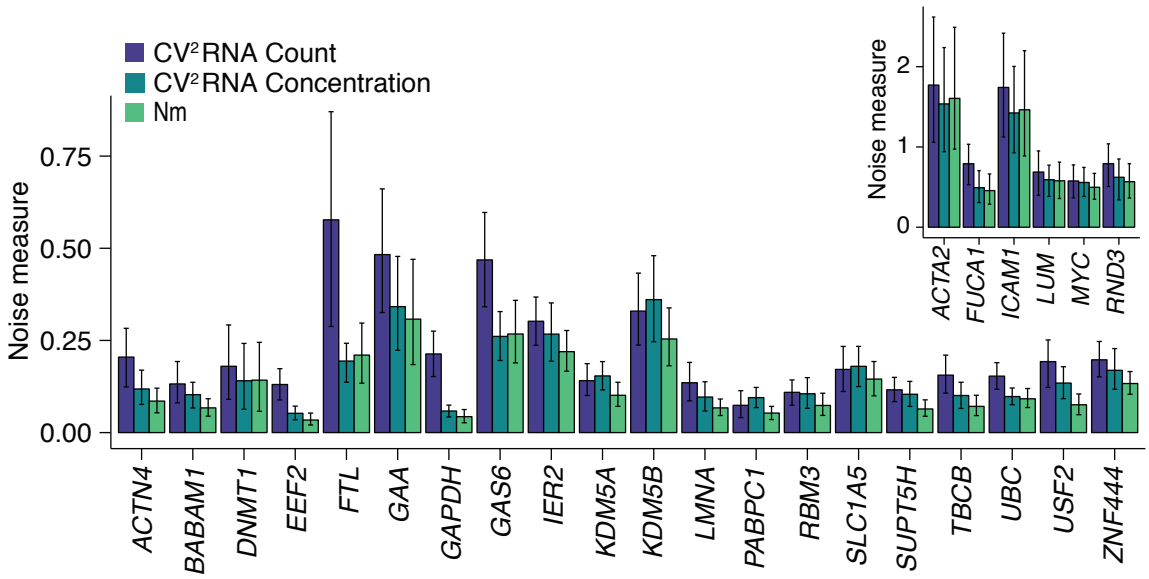

Figure 4.2: Noise measure comparison. mRNA count CV^2, mRNA concentration CV^2, and volume-corrected noise measure for cycling primary fibroblast cells. Inset shows genes that exhibit higher cell-to-cell variability in RNA, and had values too high for main axes. Generally, mRNA CV^2 is highest, followed by concentration CV^2 and volume-corrected noise measure. Error bars represent 95% confidence intervals by bootstrapping.

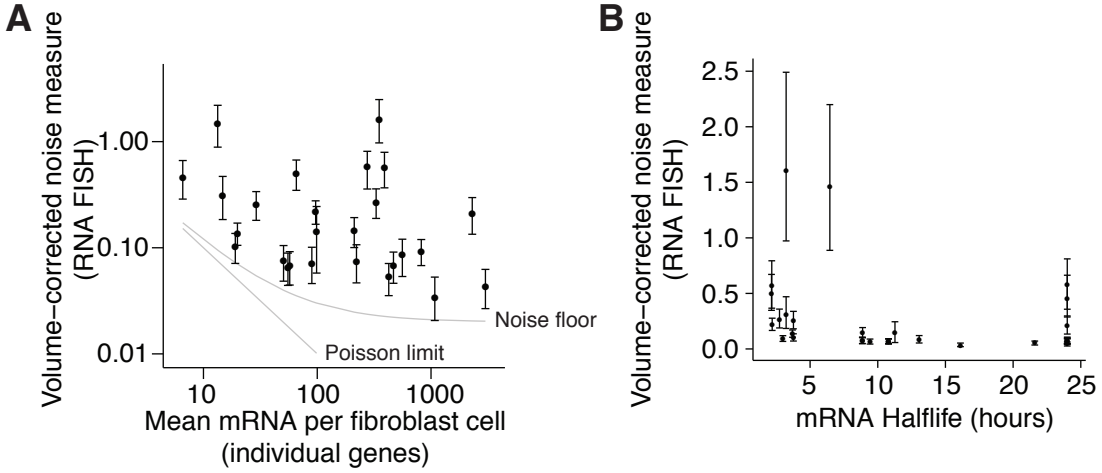

Figure 4.3: Volume-corrected noise measure does not depend strongly on mRNA abundance or half-life. (A) Volume-corrected noise measure values for different genes in primary fibroblast cells. Each data point represents a collection of single-cell measurements for one gene. The straight gray line represents the Poisson limit. The curved gray line is the Poisson limit plus our experimental noise limit, a combination of the Poisson limit and a 15% measurement error. Error bars represent 95% confidence interval by bootstrapping. Data for each gene is a combination of at least two biological replicates, with at least 30 cells per replicate. (B) We compared volume-corrected noise measure and mRNA half-life. We obtained half-life values from [62]. Volume-corrected noise measure and mRNA half-life. Each data point represents one gene. For each gene, we have at least two biological replicates with at least 30 cells per replicate. Error bars represent 95% confidence intervals, calculated by bootstrapping.

30 genes, allowing us to see that different genes show different noise levels. However, 30 data points are insufficient to draw large-scale inferences about noise profiles of various classes of genes, so we were interested in calculating variability genome-wide. To quantify this variability for all genes in the genome, we performed single-cell RNA sequencing [6, 21, 57] on human foreskin fibroblast cells, and calibrated the data such that we were able to extract volume and mRNA counts for all genes in 44 different single cells.

4.3 Calibration of single-cell sequencing data to RNA FISH

We prepared single cells for RNA sequencing using the Fluidigm C1 Single-Cell Auto Prep System, a microfluidic device that transfers single live cells into individual wells, lyses each cell and performs first- and second-strand synthesis within the device. After processing the RNA in the device, we extracted cDNA and prepared libraries for RNA sequencing using the Nextera XT library preparation kit, and sequenced on an Illumina NextSeq 500. In order to extract cell size and calibrate our RNA sequencing data to our FISH data (Fig. 4.4), we added synthetic RNA from the External RNA Controls Consortium (ERCCs [14]) at known concentrations to the C1 device along with our live cells, and sequenced both genomic RNA and ERCC RNA simultaneously. We aligned reads to the hg19 build of the human genome with annotation added for the ERCC control RNA sequences using STAR [15], and extracted counts per gene using HTSeq [1]. To calculate FPKM (fragments per kilobase per million fragments mapped), we manually extracted the "maximum exon length"—the length of the union of all annotated exons for each gene—and normalized the number of reads mapping to each gene by that gene's maximum exon length (divided by 1000) and the number

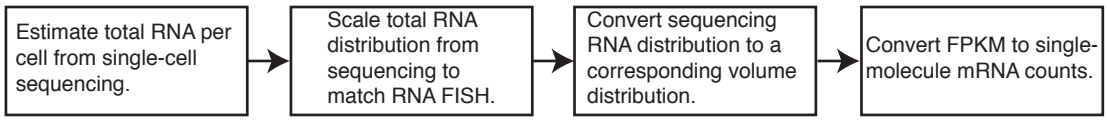

Figure 4.4: Summary of single-cell RNA sequencing calibration.

of total reads mapped per sample (divided by 10^6).

We assumed that each well received an equal amount of the ERCC control RNA, and estimated the total amount of RNA in each cell by comparing the number of reads mapping to the transcriptome for each cell to the number of reads mapping to the ERCC control RNA [36, 74]. We assumed that this ratio of (genomic reads : ERCC reads) was an accurate relative measure of total RNA per cell. There were 21 cells that had an extremely low genomic:ERCC ratio, which we assumed to be due to library preparation or sequencing artifacts, and we excluded these cells from our analysis (Fig. 4.5).

In addition to allowing us to estimate total RNA content per cell, the ERCC control RNAs give us an idea of a reliable FPKM cutoff—the FPKM value below which there is no longer a linear relationship between FPKM and the actual amount of RNA in the cell. The ERCC control RNAs are a mix of synthetic RNAs of different lengths at various relative concentrations. We mapped FPKM to ERCC RNA concentration, and found that below approximately 10 FPKM, FPKM was no longer predictive of concentration (Fig. 4.6A). Therefore in our further analyses, we only examined genomic RNAs with FPKM > 10 to ensure that the values we used were an accurate representation of the abundance of that RNA in the cell.

We have observed through RNA FISH that *GAPDH* mRNA is highly abundant and strongly correlated with cellular volume, and we therefore assumed *GAPDH* mRNA to be proportional to the total amount of mRNA in the cell. From here, we found a

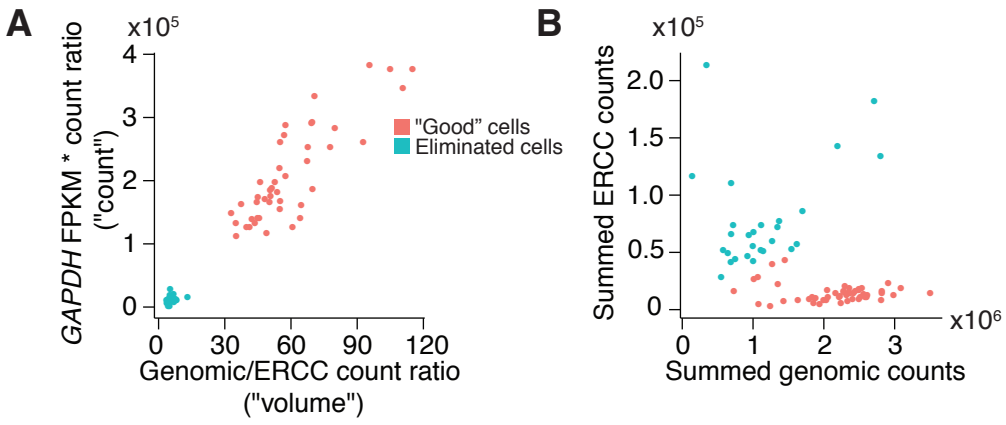

Figure 4.5: We eliminated low-"volume" cells from our analysis. (A) "Count" vs. "volume" for *GAPDH* from single-cell sequencing data. We define "volume" as the ratio between genomic reads and ERCC reads for each cell. This quantity is more representative of total RNA, which we know to be roughly proportional to volume, although the relationship is not exactly proportional due to volume-independent transcription (see Fig 2.9). We observed two clearly distinct classes of cells, those with a volume range that matches what we see by imaging and RNA FISH and those that have very low volumes. For unknown reasons, these cells ended up with a considerably higher ratio of ERCC reads than genomic reads, and we eliminated them from our subsequent analyses. (B) ERCC counts vs. genomic counts for the cells that we kept and those we eliminated.

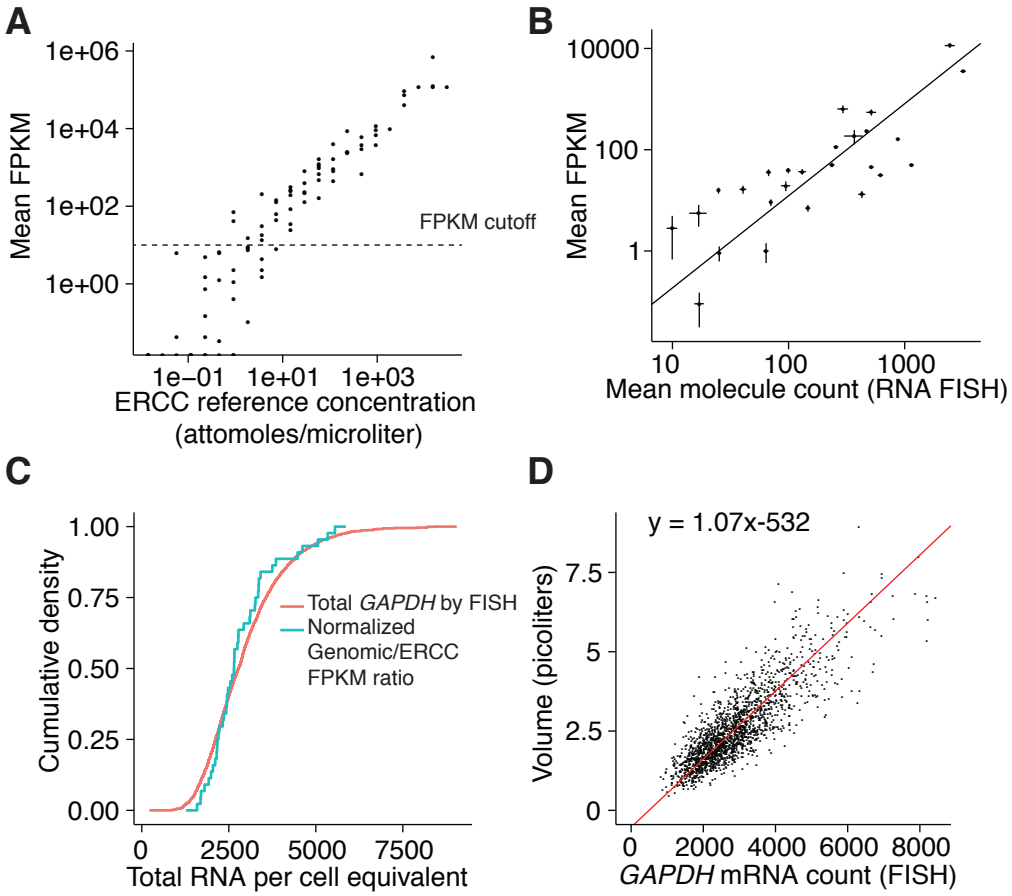

Figure 4.6: Calibration of single-cell RNA sequencing data. (A) Mean FPKM and known concentrations for each of the ERCC reference transcripts. Each point represents a single ERCC transcript and is an average over all 96 samples. An FPKM of 10 is our cutoff for "reliable" measurements. (B) Mean count as measured by RNA FISH vs. mean FPKM from single-cell RNA sequencing. Each point represents a single gene and is an average over 44 single cells for single-cell sequencing, and an average over at least two biological replicates with at least 30 cells apiece for RNA FISH. Error bars represent standard error of the mean. (C) Comparison of "total RNA" distributions from single-cell sequencing and RNA FISH. We assume that total *GAPDH* mRNA counts by RNA FISH are proportional to total RNA. For sequencing data, we use the ratio of genomic reads to ERCC reads as a proxy for total RNA. We scaled this ratio to have the same mean as the distribution of total RNA by RNA FISH. (D) Mapping between total RNA count (here, total *GAPDH* mRNA in single cells) and volume, as measured by RNA FISH. Each point represents a single cell. We use this mapping to convert total RNA from sequencing experiments to actual volume. The red line is the best fit, as computed by principle components analysis.

linear relationship between "total RNA" and cellular volume (Fig. 4.6D). In using a linear regression to find this relationship, the best fit line changes depending on which variable is considered as the dependent variable. Instead of using ordinary least squares, we used orthogonal regression, or total least squares, which uses principal component analysis to find the direction of maximum variance in a given dataset. For two variables that are linearly dependent, the first principal component picks out the direction of a best fit line that is invariant to choice of dependent variable.

We normalized our genomic:ERCC "total RNA" measurements to have the same mean as the distribution of total *GAPDH* mRNA found through RNA FISH, and compared the two distributions (Fig 4.6C). The two distributions of "total RNA" found through RNA FISH and sequencing were remarkably similar, and we therefore used our relationship between "total RNA" and volume by RNA FISH to estimate the volume of the individual cells that we sequenced.

From our RNA FISH data, we found a linear relationship between "total RNA" (total *GAPDH* RNA) and volume, and we used this relationship to convert total RNA to relative cellular volume for the cells we sequenced. We then used the correlation between FPKM and RNA FISH counts for our gene panel (Fig. 4.6B) to provide estimates of absolute RNA counts for all genes in individual cells. It is important to note, however, that there is substantial variance in the relationship between FPKM and FISH count, and that the relationship is nonlinear. The relationship between FPKM and FISH count is best fit linearly on a log-log plot, leaving us with the following nonlinear relationship:

$$FISHcount = 10^{(\log(FPKM)-a)/b},$$

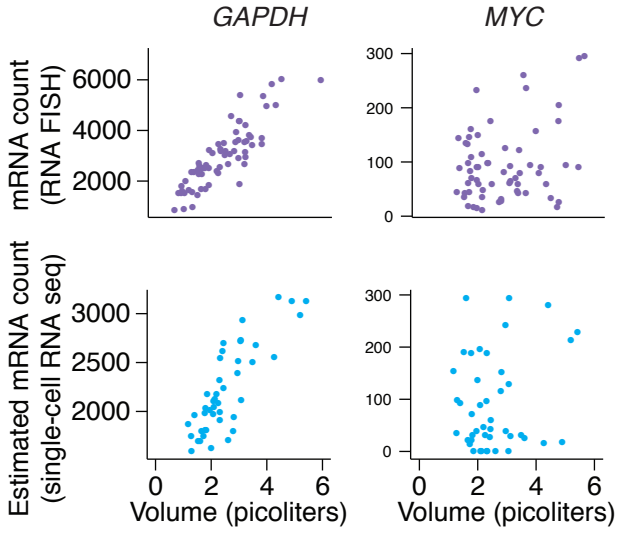

Figure 4.7: Qualitative comparison of count vs. volume from RNA FISH and single-cell RNA sequencing. Example low-Nm ($GAPDH$) and high-Nm gene (MYC).

where a and b are the intercept and slope of the best-fit line on the log-log plot.

Finally, by assigning volumes to each of the cells in our single-cell sequencing experiments and estimating RNA abundance in terms of absolute counts instead of FPKM, we reproduced our RNA vs. volume plots using RNA sequencing data (Fig. 4.7), and found that qualitative trends we observed by RNA FISH were also present in the sequencing data. From this data, we calculated Nm for every gene in the genome that had a mean FPKM of 10 or greater. We found that Nm calculated from our single-cell RNA sequencing data was a good estimate of Nm calculated by RNA FISH (Fig. 4.8).

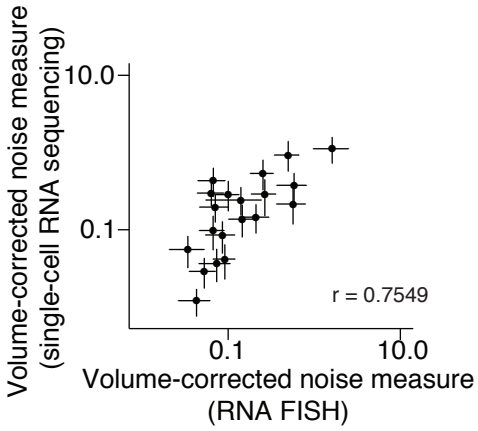

Figure 4.8: Comparison between *Nm* calculated from RNA FISH data and single-cell RNA-seq data. Each point represents a single gene. *Nm* is calculated by bootstrapping; error bars represent 95% confidence interval, calculated by bootstrapping.

4.4 Cell-type specific genes are noisier than ubiquitously-expressed genes

Using our RNA FISH data, we quantified *Nm* for genes in both primary human fibroblast cells and lung cancer cells. In comparing both *Nm* and RNA count in these two cell types, we noticed that three of the four genes (*ICAM1*, *LUM*, *ACTA2*) with the strongest degree of cell-type expression specificity were also the three genes with the highest noise measure of all the genes in our study (Fig. 4.9). To see if this trend held more generally, we used our single cell RNA-sequencing data to explore noise measure across all genes.

We had bulk RNA sequencing data for our primary fibroblast cells as well as for our lung cancer cell line. Comparing FPKM across the two cell lines, we classified genes as being ubiquitously-expressed (FPKM values in both cell lines are within 2 FPKM of each other) or cell type-specific (FPKM value in one cell line is at least five times greater than in the other) (Fig. 4.10). We also classified genes as having

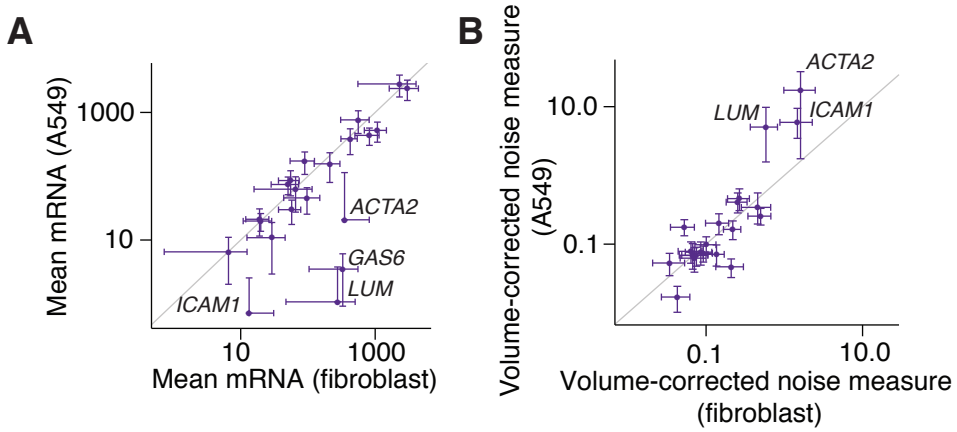

Figure 4.9: Abundance and Nm comparison in fibroblast and A549 cells. (A) Average mRNA counts in cycling primary fibroblasts and A549 cells, calculated using RNA FISH. Gray line indicates a 1:1 correspondence. Error bars represent standard error of the mean. (B) Volume-corrected noise measure in cycling primary fibroblast and A549 cells, calculated using RNA FISH. Gray line indicates a 1:1 correspondence. Nm calculated by bootstrapping; error bars represent 95% confidence interval. Data for each gene are a combination of at least two biological replicates, with at least 30 cells per replicate.

high or low noise measures, using the Nm values we calculated for all genes using our calibrated single-cell RNA sequencing data. On average, genes with higher abundances had lower noise measures, so we wanted to choose genes that had abnormally high or low noise measures for a given abundance. To do this, we transformed our Nm values such that the best-fit line between log(Nm) and log(FPKM) had a slope of zero. We then considered genes with a log(transformed Nm) > 0.5 to be high noise genes and genes with a log(transformed Nm) <-0.5 to be low noise genes (Fig. 4.10).

We selected genes with high or low noise measures and looked for enrichment in genes exhibiting cell-type specific expression between human lung cancer cell and fibroblast cell data. We found that the set of high noise measure genes contained a significantly higher proportion of cell-type specific genes than low-noise genes (Fig. 4.11). Such findings mirror those showing that more ubiquitously expressed "housekeeping" genes typically exhibit lower levels of noise than other types of genes, although the notion of cell-type specificity is more difficult to relate to studies performed in single-celled organisms [2, 42, 63].

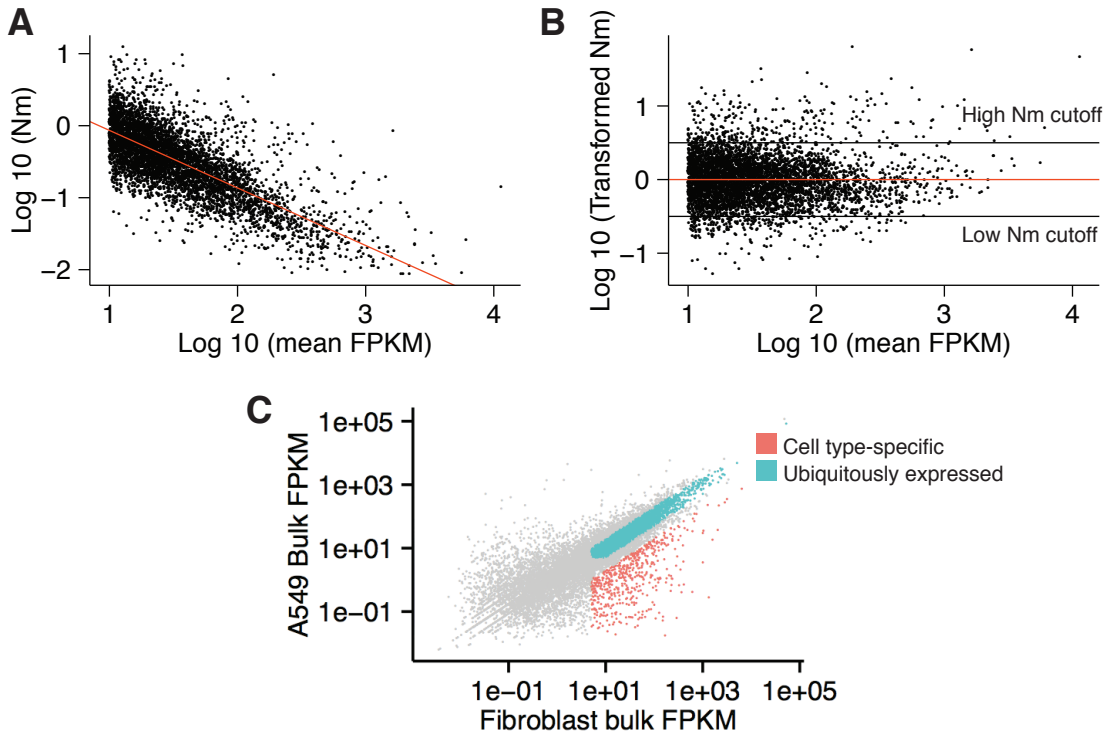

Figure 4.10: Classification of noisy and cell type-specific genes. (A) Volume-corrected noise measures from single-cell RNA sequencing data in primary fibroblast cells. Each point represents one gene. We observe that higher abundance genes typically have lower Nm values. Red line indicates best fit line. (B) The same data as in A, but transformed to remove the abundance dependence from Nm. Red line here is the transformed fit line from A. We use this transformed data to select abundance-matched "low Nm" and "high Nm" genes using a cutoff of Nm=0.5 and Nm=-0.5, respectively. We selected 307 high Nm genes and 257 low Nm genes. Note that these high Nm genes actually have a higher mean abundance (FPKM=196.5) than the low Nm genes (FPKM=55.4), thus showing that the observed differences in noise levels are not due to the overall increase in noise in genes of low abundance. (C) FPKM measurements from bulk RNA sequencing in primary fibroblast and A549 cells. Each point represents one gene. We classified genes as "ubiquitously expressed" if they had >5 FPKM in both cell types and differed by less than a factor of 2 in FPKM across the two cell types. We considered genes "fibroblast specific" if they had >5 FPKM in fibroblasts and their FPKM was greater than five times higher in fibroblasts that A549 cells.

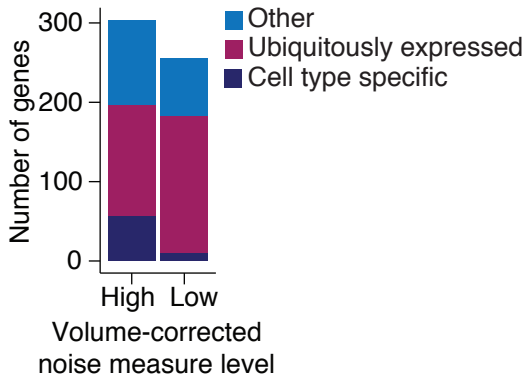

Figure 4.11: High noise genes are enriched for cell-type specific genes. Breakdown of high- and low-noise genes into ubiquitously-expressed genes and genes that express in a cell-type-dependent manner. See Fig. 4.10 for classification criteria.

Chapter 5

Discussion

We have shown that, for many genes, RNA count correlates strongly with volume, regardless of position in the cell cycle and the number of DNA copies in the cell. Moreover, we have shown that the cell employs two separate global transcriptional mechanisms to compensate for differences in volume and changes in DNA content to maintain the concentration of RNA. We expect that such generic and global mechanisms are necessary for the proper functioning and homeostasis of cells, as biological processes rely on the concentration, not the absolute count, of biomolecules within the cell. This is important in a number of biological contexts such as development and embryogenesis, in which rapid cell divisions lead to an exponential decrease in individual cell volume, but the organism must maintain the concentration of most proteins while still enabling dynamic transcriptional programs to occur [41].

Cells compensate for changes in cellular volume by increasing transcriptional burst size in larger cells. Essentially, the same gene in a larger cell produces more RNA every time it turns ON than it would in a smaller cell. How might this happen on the molecular level? We have shown that there is overall more transcriptional machinery in larger cells than in smaller cells, and that the size of transcriptional bursts is reduced when the amount of transcriptional machinery is reduced. We speculate, therefore, that the factor linking volume and transcription is the RNA polymerase II holoenzyme or another component of the general transcriptional machinery. Thus, it is likely

that larger burst size is a result of a longer "polymerase train" more polymerases simultaneously moving down the body of a gene, producing more mRNA in larger cells. It could be the case that either (a) polymerases pile up at the promoter when the gene is in an OFF state, and all traverse the length of the gene once it turns ON, or (b) once a gene turns ON, diffusing polymerases interact with the promoter and the transcription initiation complex, stochastically initiating transcription. Either scenario would give us bigger transcriptional bursts in larger cells, simply because there is more polymerase in the nuclei of larger cells. Other studies [34, 79] have speculated that the RNA polymerase II holoenzyme may act as a limiting factor titrated by DNA, but various studies [9, 26, 27] show that RNA polymerase II is only directly associated with DNA for short periods of time. Thus, it is perhaps more likely that scenario (b) holds, although we cannot rule out either possibility with our current data.

We found it striking that the volume compensation mechanism is distinct from the one that compensates for changes in DNA content as the cell cycle progresses. We found that the burst frequency appears to decrease upon DNA replication for each gene rather than at a particular time in the cell cycle for all genes. One plausible explanation for this feature being separate from the volume compensation mechanism is to ensure proper transcriptional output regardless of whether a gene is replicated early or late in S-phase, which can proceed for many hours. The molecular underpinnings of this mechanism remain unclear, although our results demonstrate that it must be a factor that remains bound to DNA and changes in character during DNA replication. A likely candidate may be a DNA or histone modification that completely coats the DNA during G1 but is diluted by a factor of two per DNA copy upon DNA replication in S-phase. The cell faces the challenge of "resetting" transcription back to G1 levels after cell division, and we suspect that the cell accomplishes this by modifying DNA or histones during M phase [68].

The two mechanisms described here allow RNA production to scale with volume, helping to maintain the concentration of mRNA between cells. However, we have shown that mRNA concentration is in fact not constant between cells for two different reasons. First, there are many genes with very "noisy" mRNA expression, whose mRNA does not exhibit any correlation with volume. Second, even genes whose RNA expression is tightly correlated with volume display higher mRNA concentrations in small cells than large cells, as a result of volume-independent transcription.

To the second point, for many genes, we observe significantly different mRNA concentrations in the smallest and largest cells, sometimes by a factor of two or more. Overall, concentration is less variable than mRNA count, but it is far from constant. Why might this be the case? We observe that smaller cells have a higher concentration of mRNA than larger cells, leading to a positive non-zero intercept for the fit line between mRNA and volume. It may be that small cells produce higher concentrations of RNA and protein as a catalyst for growth, although we do not have sufficient data to test this hypothesis. Interestingly, we observe that overall, volume-independent expression is lower in growth-arrested cells than in cycling cells, providing more evidence that there may be a link between volume-indpendent transcription and growth.

How, mechanistically, do smaller cells have a higher concentration of mRNA than larger cells? In our model for the volume/DNA sensor that links transcription and volume, a factor is expressed proportional to cellular volume, primarily localized to the nucleus, and transcriptional activity is dependent on the concentration of the factor in the nucleus. We observe that nuclear size increases slightly with cytoplasmic volume, so the concentration of the factor in larger cells will be slightly less than it would be if nuclear size were constant. This slight decrease in concentration in large cells may be enough to account for the mRNA concentration differences we observe in large cells

and small cells, although more experiments will be required to establish this model completely. Regardless of the origin of the effect, it is clear that mRNA concentration is typically higher in smaller cells. We do not yet have sufficient data to understand the consequences of this effect, in particular on cell growth, nor how it may vary in different cell types and contexts.

We have also observed that there are some genes whose RNA simply does not correlate with volume ($R^2 \approx 0$), and we consider these genes "noisy". Why might a gene be noisy? Genes that are turned on in response to particular stimuli, such as heat shock response genes and cell cycle-dependent genes will be less likely to scale with volume in general. If some cells have received a stimulus while others have not, we would not expect all cells to have consistent levels or concentrations of mRNA. Interestingly, for some cell cycle-dependent genes, we observe two distinct fit lines between mRNA count and volume when we gate for cell cycle. When the gene is OFF or lowly expressed, it has one characteristic mRNA concentration, and when it is ON, it has a different concentration. However, if we observe mRNA from that gene without gating for cell cycle, the data look noisy. This is likely not the case for all noisy genes, but an interesting observation nonetheless. With our single-cell RNA sequencing data, we now have the ability to look at noise levels of genes genome-wide. Using this data, we can potentially begin to make inferences about types of classes of genes that have high noise levels, and perhaps begin to study what makes genes noisy in the first place.

An interesting question is how it is even possible for genes to be highly noisy, given that there is a global transcriptional mechanism causing mRNA expression to scale with volume. In this work, we showed that transcription of the *LMNA* gene is regulated globally, not by a gene-specific network that senses *LMNA* protein concentration. However, this is not to say that genes are not also subject to individual regulation outside of the global regulation by cell volume. Rather, each gene has its own promoter

and enhancer(s), and its transcription is activated by particular transcription factors. These factors together help to determine traits such as the noise level and the average mRNA expression level of a given gene. The global mechanisms described in this work simply provide a means by which any such additional regulation may operate without having to take into account differences in DNA concentration due to cellular volume or DNA copy number differences. A consequence of this global mechanism is that the RNA from any gene, noisy or not, should scale with volume in the mean. Larger cells should have more mRNA than smaller cells, regardless of how variable the gene's expression is. Indeed, for all of the genes we measured by RNA FISH, we observe that mRNA expression is higher in the largest 25% of cells than in the smallest 25% of cells.

From our data, we conclude that transcription and volume are linked through a factor that senses both volume and DNA content. We have strong evidence that this factor is a part of the general transcriptional machinery, such as RNA polymerase II, but we have not conclusively shown this to be the case. RNA polymerase fits all of the criteria for the factor: it is expressed proportional to cellular volume, it is almost entirely nuclear, and it is important for transcription. However, we cannot rule out that there is another, more global factor that controls the expression levels of RNA polymerase. Indeed, there must be some regulation on our factor, otherwise the model as stated could lead to unregulated growth and the potential for enormous cells, which we do not observe.

In our model, larger cells have more of the factor, which leads to more transcription than in smaller cells. Similarly, because larger cells have higher levels of transcription, they are able to produce more of the factor. At steady state, the model nicely describes differences in size and transcription levels between large and small cells. However, if the system is somehow perturbed, causing a cell to produce more of the factor

than it should for its size, unrestricted growth could occur: more factor begets more transcription, which causes cells to grow, and so on. It is possible that this problem is simply solved by the fact that nuclear size scales with volume: perhaps as a cell grows, its nucleus grows in size such that transcription rate does not continually increase with volume. It is also possible that at some point, transcriptional activity can no longer increase, even with more transcriptional machinery. There may be a point at which DNA reaches its capacity, as there are only a finite number of binding sites for transcriptional machinery. Both of these possibilities would lead to a plateau or leveling-off of mRNA abundance above a certain volume. It is possible that we begin to see such an effect in our heterokaryon data, and even in large datasets of *GAPDH* mRNA in unperturbed cells. Zhurinsky and colleagues noticed this phenomenon in large yeast cells [79].

It is important to note that throughout this work we have considered cells to be static, and have not included growth in our models. It may be the case that we need to consider growth in order to have a complete understanding of how cell size and transcription are coupled. Perhaps there are independent regulators of growth that indirectly affect transcriptional activity. Further, we have not addressed the question of why cells have different volumes and how expression plays a role in that heterogeneity—such questions necessarily involve the examination of mechanisms regulating cell growth and proliferation. Rather, our results show how cells may globally cope with such changes to maintain biomolecule concentration.

The work we report here highlights the importance of taking cellular volume into account when interpreting gene expression data and points to the significance of global factors in studying single cell expression in general [13, 16, 70]. In particular, our cell fusion experiments show that changing the amount of cellular content in and of itself can lead to changes in total RNA abundance, whereas previous experiments largely relied on

cell-size mutants that make inferences of cause and effect more difficult [19, 39, 55, 79]. These cell fusion experiments directly establish that any perturbation that changes cellular volume may result in global changes in overall transcript abundance as a secondary rather than primary effect per the generic mechanism of a diffusible *trans* factor that senses an increased ratio of volume to DNA. Thus, we believe one must take care in interpreting experiments showing global changes in transcript abundance [32, 43], both from the perspective of establishing causal relationships, given that cellular volume/content can by itself change transcription rates, and in the interpretation of the functional significance, given that the concentration of many transcripts will remain roughly the same despite these overall changes. An example of this is our comparison of *GAPDH* mRNA levels in cycling and growth-arrested cells. Cycling cells, on average, had more *GAPDH* mRNA molecules than the growth-arrested cells. However, the concentration of *GAPDH* mRNA between the two types of cells was approximately the same, due to the fact that growth arrested cells were smaller on average than the cycling cells. Here, by simply counting mRNA, it would appear that cycling cells are more active, when really the cells are just larger. Measuring mRNA concentration instead of simply counting mRNA will resolve such interpretation issues.

On the one hand, we have shown that it is important to take volume into consideration, and that mRNA concentration is a better metric in many ways than mRNA count. However, we have also shown that mRNA concentration is not constant between cells. One practical consequence of this finding is that the time-honored practice of normalizing transcript data, be it qRT-PCR or RNA sequencing or RNA FISH, to *GAPDH* mRNA abundance, while largely sound, does not fully account for differences in mRNA concentration between small and large cells. This suggests that new strategies may be required for measuring cellular volume when interpreting PCR or sequencing data, particularly in single cell RNA-sequencing experiments.

Together, these findings provide a deeper quantitative understanding of single cell gene expression and its role in maintaining cellular homeostasis. Further work may elucidate how these homeostatic mechanisms for maintaining biomolecule concentration manifest themselves in biological contexts and whether they are an important point of dysregulation in disease processes.

Appendix A

Experimental and computational methods

Cell culture

We grew primary human foreskin fibroblast cells (CCD-1079Sk, ATCC CRL-2097^{TM}) and A549 cells (human lung carcinoma, A549, ATCC CCL-185^{TM}) in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% FBS and 50U/mL penicillin and streptomycin (Pen/Strep). To create quiescent cells, we grew primary fibroblast cells in DMEM with Pen/Strep, without FBS for seven days. We cultured WM983b-GFP-NLS cells (WM983b is a human melanoma cell line from the lab of Meenhard Herlyn) in Tu2% media (78% MCDB media, 20% Leibovitz's L-15 media, 2% FBS, and 1.68 mM CaCl2). The WM983b-GFP-NLS contains EGFP fused to a NLS driven by a cytomegalovirus promoter that we stably transfected into the parental cell line. Before imaging, we plated cells on two-well Lab-Tek chambered coverglasses.

RNA fluorescence *in situ* hybridization and imaging

We performed single molecule RNA FISH on the samples as described previously [18, 48, 49]. Briefly, we fixed the cells in formaldehyde or methanol, performed RNA FISH using the specified pools of oligonucleotides, then washed and stained nuclei with DAPI. We fixed most cells in this study using formaldehyde, but used methanol for the experiments involving transcription site quantification because it resulted in more accurate transcription site detection. We stained the actin cytoskeleton with Phalloidin-Alexa 488 (Life Technologies) to detect cell boundaries.

We typically co-stained with sets of probes targeting many different mRNA. Typically, we used exon probes labeled in Alexa 594, introns with Cy3, Cyclin A2 with Atto 647N (which labels cells in S, G2 and M phase [17]) and *GAPDH* with Atto 700. To distinguish cells in S phase from G2, we labeled Histone H4 mRNA with Atto 647N and Cyclin A2 mRNA with Atto 700 [53, 73].

We imaged the cells with a Nikon Ti-E equipped with appropriate filter sets. We took a series of optical z-sections, each 0.2-0.35 microns high, that spanned the vertical extent of the cell.

Image analysis and quantification

We manually identified cell boundaries and counted and localized RNA spots using custom software written in MATLAB [48, 49]. We estimate the technical error in our RNA count determination to be at most 15%.

To compute the volume of a cell, we detected the 3D positions of a highly abundant mRNA by RNA FISH. We selected only the points that defined the outer boundary

of the cell by examining each point and its neighbors within a 4μm radius. We kept only the points that had a higher z-position than their neighbors (signifying the top of the cell) or points that had no neighbors within 180 degrees (signifying the side of the cell). Once we had the points, we interpolated the points to identify a smooth representation of the cell surface. We repeated this in both an upward and downward direction to identify the top and bottom of the cell. We calculated the volume of the cell by summing the heights between the top and bottom.

Calculating the volume in this manner will always result in an underestimation of the actual volume. To correct for this bias, we first computed the outline of the cell as described above. We then dilated this hull, filled it with the same number of randomly distributed points, and then repeated the algorithm on this new set of points. If this volume matched that computed with the actual spots in the cell, we then computed the volume by integrating between the top and bottom boundaries of the dilated hull.

We used *GAPDH* mRNA as the primary mRNA for our volume determinations, but the volume computation did not depend on the number of spots identified nor on the choice of volume-filling gene. We limited ourselves to the cytoplasmic volume by removing a vertical cylinder corresponding to the nuclear outline. This procedure does exclude the cytoplasmic volume above and below the nucleus, but that region only comprised a very small proportion of the total cytoplasmic volume.

We identified transcription sites through intron/exon probe colocalization. We manually annotated transcription sites by visually inspecting images of intron and exon probes to determine instances of colocalized signal. To determine spot intensity, we identified the z-plane of maximum intensity in a 0.375μm-square region around the manually selected spot. We defined the intensity as the difference between this maximum value and a background value. For the background value, we used the median intensity in a 3.75μm-square annular region around the maximum intensity

point. Note that transcription site intensity need not necessarily linearly relate to transcriptional burst size [56].

RNA degradation

We measured RNA degradation by inhibiting transcription for four hours by applying actinomycin D at 1μg/ml. We measured degradation of *UBC* and *IER2* mRNA because they exhibited a strong correlation with volume while having a half-life short enough to enable us to observe substantial degradation within four hours of actinomycin D treatment while avoiding non-specific effects at longer times.

We used a model to determine whether degradation was volume-dependent (degradation $\sim 1/V$) or volume-independent (degradation $\sim$ constant). We first fit the untreated mRNA vs. volume data with a line having zero intercept. If degradation is volume-independent, we expect the treated cells to also be well-fit by a line having zero intercept, where the slope is determined by the untreated fit and an exponential decay term:

$$m_{4h}(V) = s_0 V e^{-\gamma t},$$

where m_{4h} is the mRNA count after 4 hours of treatment, s_0 is the slope of the untreated data (t=0), γ is the decay constant (degradation rate), and t is the treatment time. Note that here γ is the only fit parameter and is independent of volume.

If degradation is volume-dependent, the equation becomes:

$$m_{4h}(V) = s_0 V e^{-\gamma t/V}.$$

Here, γ/V is the decay constant (degradation rate), but γ itself is independent of volume and is the fit parameter.

The line and curve described by these equations are the fits to the raw data, and the decay constants γ and γ/V are the fits we show to the calculated decay constants that we show in Fig. 2.11.

We calculated the actual decay constant for each cell measured at the 4 hour timepoint assuming exponential decay:

$$m_{4h}(V) = m_{0h}(V)e^{-\gamma t},$$

where we approximate $m_{0h} = s_0 V$, and γ could in principle be either volume-dependent or -independent.

LMNA siRNA knockdown

We used an siRNA targeting LMNA (Cat. #: AM16708, ID: 40502) at 30nM and a "scramble" control siRNA (Cat. #: AM4611) at 30nM. We incubated primary fibroblast cells with the siRNA for 72 hours. We verified protein knockdown via Western blot, using the SC-20680 (rabbit) antibody and a goat-anti-rabbit 680 RD secondary (Licor 926-68071).

Heterokaryon formation

We created heterokaryons by separately culturing primary fibroblast cells and WM983b-GFP-NLS cells. Once the plates were 70-90% confluent, we trypsinized the cells, resuspended them in DMEM Complete media, and combined half of each plate of cells in a 15ml tube. We pelleted the cells and resuspended in PEG for 2 minutes. We added media over the course of five minutes to allow cells to fuse, then plated the cells onto two-well chambered coverglasses (Lab-Tek) and fixed the cells after 12

hours.

We identified heterokaryons as cells with two nuclei that expressed both GFP (WM983b-GFP-NLS only) and *GAS6* mRNA (primary fibroblast) by RNA FISH. We eliminated all homokaryons (two cells of one type fused together) from our analyses.

Fractionation and RNA polymerase II Western blot

This protocol was performed by the Churchman lab at Harvard. We performed cell fractionation as described in [4] and based on [76] with modifications. We conducted all subsequent steps on ice or at 4°C and in the presence of 25 μM α-amanitin (Sigma, A2263) and Protease inhibitors cOmplete (Roche, 11873580001) according to manufacturer's instructions. We pre-chilled all buffers on ice before use. We grew primary fibroblast cells to a confluency of 90%. We removed media and washed plates twice with 1x PBS before scraping cells into 1x PBS. We collected cells by centrifuging at 500 g for 10 min. We gently resuspended the cell pellet corresponding to 1×10^7 cells in 200 μl cytoplasmic lysis buffer (0.15% NP-40, 10 mM Tris-HCl pH 7.0, 150 mM NaCl). We incubated the cell lysate for 5 min on ice, layered it onto 500 μl sucrose buffer (10 mM Tris-HCl pH 7.0, 150 mM NaCl, 25% sucrose) and centrifuged at 16,000 g for 10 min. We carefully removed the supernatant (600 μl) corresponding to the cytoplasmic fraction. We gently resuspended the nuclei pellet in 400 μl nuclei wash buffer (0.1% Triton-X-100, 1 mM EDTA, in 1x PBS) and centrifuged it at 1,500 g for 1 min. We removed the supernatant and gently resuspended the pellet in 200 μl glycerol buffer (20 mM Tris-HCl pH 8.0, 75 mM NaCl, 0.5 mM EDTA, 50% glycerol, 0.85 mM DTT). Next, we added 200 μl nuclei lysis buffer (1% NP-40, 20 mM Hepes pH 7.5, 300 mM NaCl, 1M Urea, 0.2 mM EDTA, 1 mM DTT), vortexed, incubated on ice for 2 min and centrifuged at 18,500 g for 2 min. We carefully removed the supernatant corresponding

to the nucleoplasmic fraction (350 μl) and added 250 μl 1x PBS/Protease inhibitors cOmplete to adjust the volume for Western blot experiments (described below). We resuspended the chromatin pellet in 600 μl chromatin resuspension solution (25 μM α-amanitin, Protease inhibitors cOmplete, in 1x PBS).

We monitored the success of cell fractionation by Western blot analyses. For Western blot analyses, we probed membranes with the following primary antibodies: Pol II (F-12, Santa Cruz Biotechnology; directed against the N-terminal region of Rpb1), Pol II Ser2-P (3E10, Active Motif), Pol II Ser5-P (3E8, Active Motif), Histone 2B (FL-126, Santa Cruz Biotechnology), U1 snRNP70 (C-18, Santa Cruz Biotechnology) and *GAPDH* (6C5, Applied Biosystems). Next, we probed membranes with Cy5- and Alexa Fluor 647-conjugated secondary antibodies (Cy5 goat anti-mouse, A10524; Cy5 goat anti-rabbit, A10523; Cy5 goat anti-rat, A10525; Alexa Fluor 647 rabbit anti-goat, A21446; Life Technologies), and scanned using a Typhoon 9400 scanner (GE Healthcare). We quantified fluorescent signals with ImageJ 1.47v software.

Triptolide

We degraded RNA polymerase II in primary fibroblast cells by incubating cells in 100nM triptolide for one hour, then fixed cells in methanol (control cells remained untreated).

Cell size verification

To check that fixation did not alter cell size, we monitored the size of cells through the fixation and permeabilization process by fixing cells while on the microscope stage. We monitored cell area by taking images in brightfield, and we monitored cell height by coating the cells with fluorescent beads and imaging them in a series of

optical z-sections. We took images of the same cells after 10 minutes of fixing in 4% formaldehyde and after 30 minutes of permeabilization in 70% ethanol. We calculated cell area by segmenting the cells as usual, and we determined height by identifying the plane of the bottom of the cell and the plane of the top of the cell (the last plane where beads remain motionless) and subtracting the two values.

Quantification of cell-to-cell variability

We developed a phenomenological metric for cell-to-cell variability that takes into account both volume-correlated and volume-independent contributions to mRNA numbers per cell (see Section 4.1 for derivation and further information). We also used a model of transcriptional bursting with volume-dependent transcription that enabled us to quantify transcriptional parameters from population distributions of mRNA counts and volumes.

Repli-seq analysis

We accessed Repli-seq data from Hansen et al. [22] using the UW Repli-seq track on the UCSC Genome Browser.

Bulk RNA Sequencing

We sequenced total RNA from primary fibroblast cells. We used the NEB Next Ultimate Library Preparation Kit for Illumina and the Ribo-Zero Magnetic Gold Kit. We used 50b paired-end reads and sequenced each of two replicates at a depth of 10-15M reads. We aligned reads to hg19 using STAR's included annotation [15]. We quantified reads per gene using HTSeq [1] and a RefSeq hg19 annotation. We

calculated FPKM for each gene using R. All sequencing data is available at GEO accession number GSE66053.

Single-cell RNA Sequencing

We isolated 96 single cells, lysed, and performed first- and second-strand synthesis on a Fluidigm C1 Single-Cell Auto Prep System using a large size chip. We spiked in ERCC (External RNA Controls Consortium) RNA controls, Mix 1 (Ambion 4456740) at a concentration of 1:10,000 before adding the cells to the C1. We prepared cDNA libraries using the Nextera XT DNA Sample Preparation Kit (Illumina, PN FC-131-1096) and used 96 paired barcodes from the Nextera XT DNA Sample Preparation Index Kit (96 Indices, 385 Samples) (Illumina, PN FC-131-1002) following the abbreviated Fluidigm protocol for the Nextera XT kit. We sequenced the samples on a NextSeq 500 using 75b paired-end reads to a depth of ∼1-2M reads per sample. To quantify sequencing data, we aligned reads to hg19 (using STAR's included annotation) and the ERCC reference transcripts. We quantified reads per gene using HTSeq and a RefSeq hg19 annotation. All sequencing data is available at GEO accession number GSE66053.

Single-cell RNA Sequencing Calibration and Analysis

We independently calculated ERCC and genomic FPKM for each sample, normalizing to the total number of reads mapped to ERCC loci or genomic loci, respectively. All FPKM data shown for endogenous genes is this genomic FPKM. For each cell, we considered the ratio of total genomic reads to total ERCC reads to be proportional to the total starting amount of mRNA in that cell.

We sequenced 96 wells total, of which 5 were "control" wells that contained no cells and 14 were wells containing fixed cells. We excluded these 19 cells from the analysis. Further, we excluded 12 cells that had fewer than 1 million total reads, and 21 cells that had a genomic/ERCC read ratio of less than 30. We performed all further analyses on the 44 remaining cells.

Transform read ratio to volume

We assumed that the ratio of genomic/ERCC reads for each sample was proportional to the total mRNA in each cell. We also assumed that, for our RNA FISH measurements, total *GAPDH* mRNA counts were proportional to the total amount of mRNA in each cell. The distributions for total mRNA obtained in this manner were similar between RNA FISH and single-cell RNA sequencing, but had different means. We therefore normalized the sequencing data to have the same mean as the RNA FISH distribution. From our RNA FISH dataset, we have many co-measurements of *GAPDH* mRNA (total mRNA) and volume from which we establish a transformation equation between total mRNA and volume. We obtained this transformation equation using PCA, or orthogonal regression. Using this equation, we transformed total mRNA obtained through sequencing into actual volume in picoliters.

Transform FPKM to molecule count

FPKM is more a measure of mRNA concentration than mRNA count, as it is normalized to total reads. To get a measure more similar to mRNA count, for each cell, we multiplied each gene's FPKM by the genomic/ERCC count ratio ("volume") of the cell. For each gene in our RNA FISH dataset, we fit the log of the seq "counts" and the log of the actual counts from RNA FISH by orthogonal regression. We then used

this transform to convert the FPKM of all genes to their RNA FISH count equivalent. Note that, because we fit in log space, the transform between FPKM and count is nonlinear, and actually scales as approximately FPKM $\sim$ (RNA FISH)$^{1.7}$.

Once we had our single-cell sequencing data in terms of RNA FISH count and volume in picoliters, we calculated Nm as described for RNA FISH. We performed all of our sequencing analysis in R.

C. elegans growth and imaging

We grew N2 (wild type) and CB502 (*sma-2* mutant) C. elegans on NGM agar plates with OP50 lawns, kept at 20° C. Every 2-3 days, we transferred a small portion of each strain to new plates to prevent overgrowth.

We released the worms off of the plates using phosphate buffered saline (PBS) solution, then fixed with 4% formaldehyde for 45 minutes. We permeabilized and stored the worms in 70% ethanol. We performed the RNA FISH protocol, then mounted the sample between a slide and coverglass before imaging.

We manually identified head and gonad boundaries and counted and localized RNA spots using custom software written in MATLAB.

To compute the volume of each worm segment, we multiplied the area of the segment by the height of the segment (thus approximating the segment as a prism). We determined the height by taking the vertical difference between the highest and lowest RNA spots' positions, as determined by our software. We determined the number of cells in each segment through manually counting the DAPI-stained nuclei.

We obtained data from multiple segments. When combining the data (number of mRNA spots per volume or per nucleus), we weighted each segment by its volume.

Appendix B

Comprehensive RNA counts and concentrations for all genes and cell types

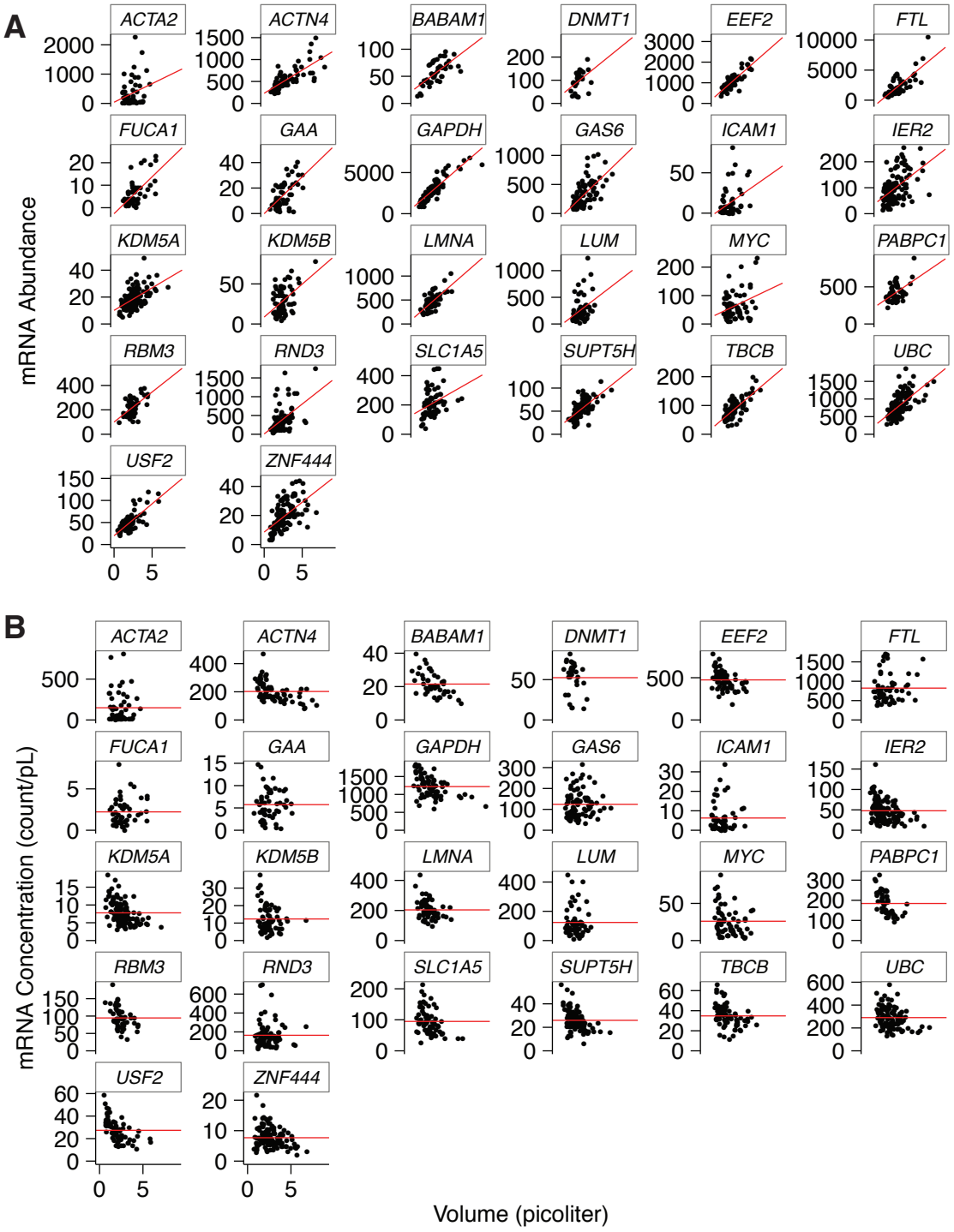

Figure B.1: Count (A) and concentration (B) of all mRNA in cycling primary human fibroblast cells. Each data point is an individual single cell measurement. In count plots, red line indicates best linear fit to the data. In concentration plots, red line indicates mean mRNA concentration. Each data set is a combination of at least two biological replicates.

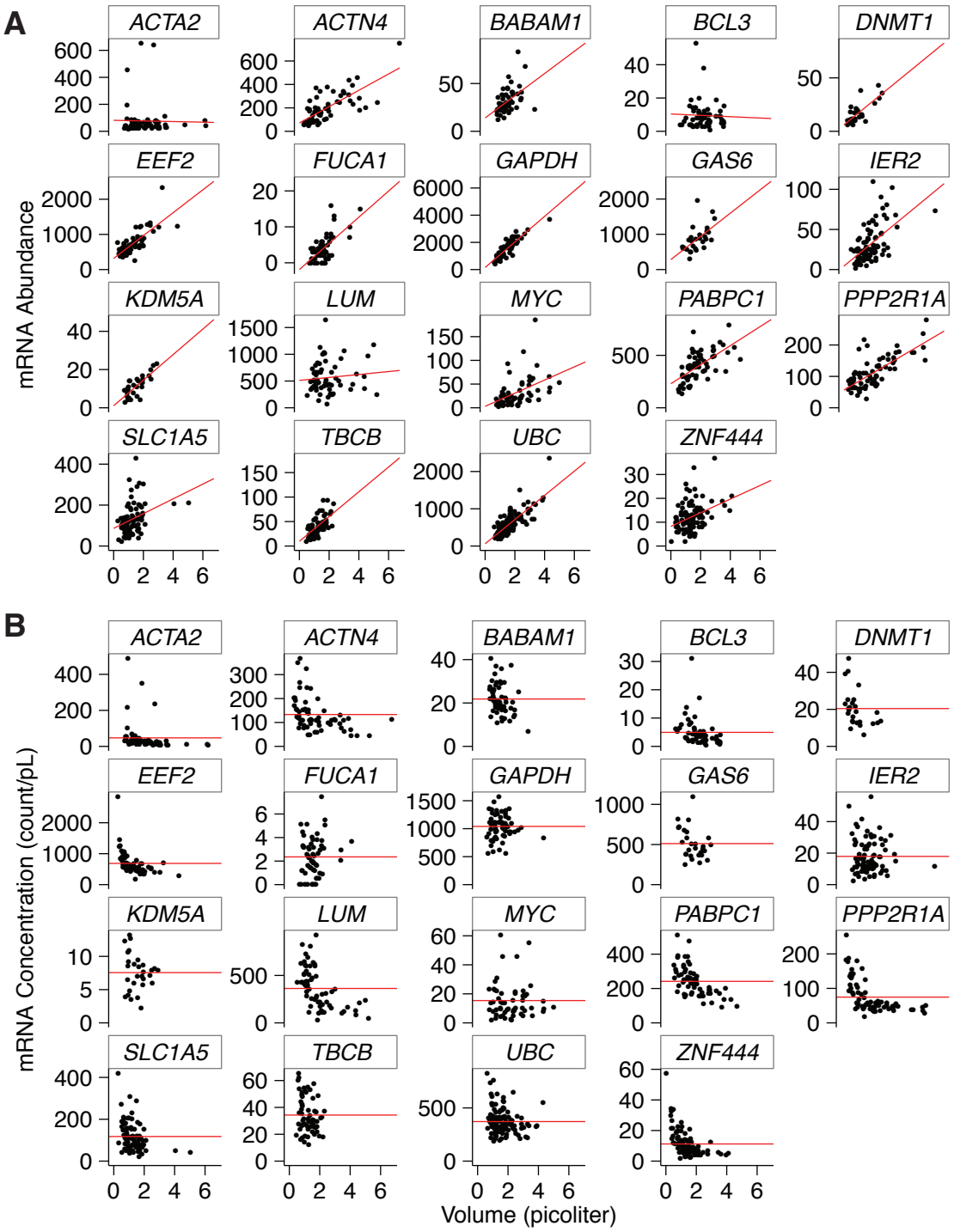

Figure B.2: Count (A) and concentration (B) of all mRNA in quiescent primary human fibroblast cells. Each data point is an individual single cell measurement. In count plots, red line indicates best linear fit to the data. In concentration plots, red line indicates mean mRNA concentration. Each data set is a combination of at least two biological replicates.

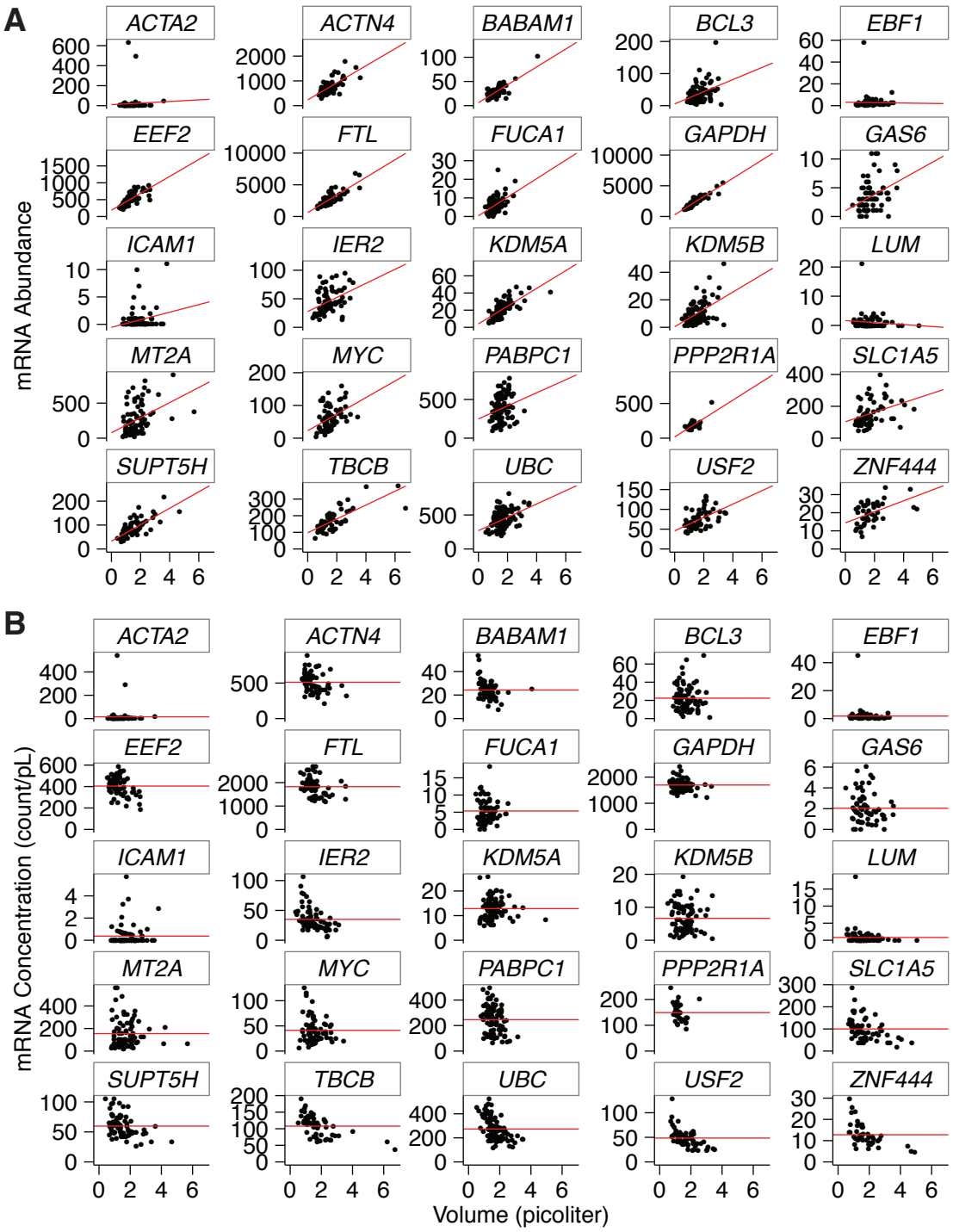

Figure B.3: Count (A) and concentration (B) of all mRNA in A549 cells. Each data point is an individual single cell measurement. In count plots, red line indicates best linear fit to the data. In concentration plots, red line indicates mean mRNA concentration. Each data set is a combination of at least two biological replicates.

Bibliography

[1] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*, 31(2):166–169, January 2015.

[2] Arren Bar-Even, Johan Paulsson, Narendra Maheshri, Miri Carmi, Erin O'Shea, Yitzhak Pilpel, and Naama Barkai. Noise in protein expression scales with natural protein abundance. *Nature Genetics*, 38(6):636–643, May 2006.

[3] Olivier Bensaude. Inhibiting eukaryotic transcription: Which compound to choose? How to evaluate its activity? *Transcription*, 2(3):103–108, May 2011.

[4] Dev M Bhatt, Amy Pandya-Jones, Ann-Jay Tong, Iros Barozzi, Michelle M Lissner, Gioacchino Natoli, Douglas L Black, and Stephen T Smale. Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell*, 150(2):279–290, July 2012.

[5] Clive G Bowsher and Peter S Swain. Identifying sources of variation and the flow of information in biochemical networks. *Proceedings of the National Academy of Sciences*, 109(20):E1320–8, May 2012.

[6] Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A Teichmann, John C Marioni, and Marcus G Heisler. Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 10(11):1093–1095, November 2013.

[7] Andrea K Bryan, Vivian C Hecht, Wenjiang Shen, Kristofor Payer, William H Grover, and Scott R Manalis. Measuring single cell mass, volume, and density with dual suspended microchannel resonators. *Lab on a chip*, 14(3):569–576, February 2014.

[8] Jonathan R Chubb, Tatjana Trcek, Shailesh M Shenoy, and Robert H Singer. Transcriptional pulsing of a developmental gene. *Current Biology*, 16(10):1018–1025, May 2006.

[9] Ibrahim I Cisse, Ignacio Izeddin, Sebastien Z Causse, Lydia Boudarene, Adrien Senecal, Leila Muresan, Claire Dugast-Darzacq, Bassam Hajj, Maxime Dahan, and Xavier Darzacq. Real-time dynamics of RNA polymerase II clustering in live human cells. *Science (New York, N.Y.)*, 341(6146):664–667, August 2013.

[10] Luca Comai. The advantages and disadvantages of being polyploid. *Nature reviews. Genetics*, 6(11):836–846, November 2005.

[11] H A Crissman and J A Steinkamp. Rapid, simultaneous measurement of DNA, protein, and cell volume in single cells from large mammalian cell populations. *The Journal of Cell Biology*, 59(3):766–771, December 1973.

[12] Roy D Dar, Brandon S Razooky, Abhyudai Singh, Thomas V Trimeloni, James M McCollum, Chris D Cox, Michael L Simpson, and Leor S Weinberger. Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proceedings of the National Academy of Sciences*, 109(43):17454–17459, October 2012.

[13] Ricardo Pires das Neves, Nick S Jones, Lorena Andreu, Rajeev Gupta, Tariq Enver, and Francisco J Iborra. Connecting variability in global transcription rate to mitochondrial variability. *PLoS Biology*, 8(12):e1000560, 2010.

[14] Alison S Devonshire, Ramnath Elaswarapu, and Carole A Foy. Evaluation of external RNA controls for the standardisation of gene expression biomarker measurements. *BMC genomics*, 11(1):662, 2010.

[15] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, 29(1):15–21, January 2013.

[16] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, August 2002.

[17] K Leigh Eward, Matthew N Van Ert, Maureen Thornton, and Charles E Helmstetter. Cyclin mRNA stability does not vary during the cell cycle. *Cell Cycle*, 3(8):1057–1061, August 2004.

[18] A M Femino, F S Fay, K Fogarty, and R H Singer. Visualization of single RNA transcripts in situ. *Science*, 280(5363):585–590, April 1998.

[19] R S Fraser and P Nurse. Altered patterns of ribonucleic acid synthesis during the cell cycle: a mechanism compensating for variation in gene concentration. *Journal of cell science*, 35:25–40, February 1979.

[20] Ido Golding, Johan Paulsson, Scott M Zawilski, and Edward C Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–1036, December 2005.

[21] Dominic Grün, Lennart Kester, and Alexander van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature Methods*, April 2014.

[22] R S Hansen, S Thomas, R Sandstrom, T K Canfield, R E Thurman, M Weaver, M O Dorschner, S M Gartler, and J A Stamatoyannopoulos. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences*, 107(1):139–144, January 2010.

[23] Takahiro Isaka, Andrea L Nestor, Tadahiro Takada, and David C Allison. Chromosomal variations within aneuploid cancer lines. *The journal of histochemistry and cytochemistry : official journal of the Histochemistry Society*, 51(10):1343–1353, October 2003.

[24] Cindy Y Jao and Adrian Salic. Exploring RNA transcription and turnover in vivo by using click chemistry. *Proceedings of the National Academy of Sciences*, 105(41):15779–15784, October 2008.

[25] Iain G Johnston, Bernadett Gaal, Ricardo Pires das Neves, Tariq Enver, Francisco J Iborra, and Nick S Jones. Mitochondrial variability as a source of extrinsic cellular noise. *PLoS Computational Biology*, 8(3):e1002416, 2012.

[26] H Kimura, Y Tao, R G Roeder, and P R Cook. Quantitation of RNA polymerase II and its transcription factors in an HeLa cell: little soluble holoenzyme but significant amounts of polymerases attached to the nuclear substructure. *Molecular and cellular biology*, 19(8):5383–5392, August 1999.

[27] Hiroshi Kimura, Kimihiko Sugaya, and Peter R Cook. The transcription cycle of RNA polymerase II in living cells. *The Journal of Cell Biology*, 159(5):777–782, December 2002.

[28] M S Ko. A stochastic model for gene induction. *Journal of Theoretical Biology*, 153(2):181–194, November 1991.

[29] Marshall J Levesque and Arjun Raj. single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. *Nature Methods*, pages 1–6, February 2013.

[30] Jeffrey M Levsky, Shailesh M Shenoy, Rossanna C Pezo, and Robert H Singer. Single-cell gene expression profiling. *Science*, 297(5582):836–840, August 2002.

[31] Gene-Wei Li and X Sunney Xie. Central dogma at the single-molecule level in living cells. *Nature*, 475(7356):308–315, July 2011.

[32] Charles Y Lin, Jakob Lovén, Peter B Rahl, Ronald M Paranal, Christopher B Burge, James E Bradner, Tong Ihn Lee, and Richard A Young. Transcriptional amplification in tumor cells with elevated c-Myc. *Cell*, 151(1):56–67, September 2012.

[33] Hédia Maamar, Moran N Cabili, John Rinn, and Arjun Raj. linc-HOXA1 is a noncoding RNA that represses Hoxa1 transcription in cis. *Genes & Development*, 27(11):1260–1271, June 2013.

[34] Samuel Marguerat and Jürg Bähler. Coordinating genome expression with cell size. *Trends in Genetics*, 28(11):560–565, November 2012.

[35] Samuel Marguerat, Alexander Schmidt, Sandra Codlin, Wei Chen, Ruedi Aebersold, and Jürg Bähler. Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell*, 151(3):671–683, October 2012.

[36] Georgi K Marinov, Brian A Williams, Ken McCue, Gary P Schroth, Jason Gertz, Richard M Myers, and Barbara J Wold. From single-cell to cell-pool transcrip-

tomes: stochasticity in gene expression and RNA splicing. *Genome Research*, 24(3):496–510, March 2014.

[37] A Mayer, J di Iulio, S Maleri, U Eser, J Vierstra, A Reynolds, R Sandstrom, J A Stamatoyannopoulos, and L S Churchman. Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell*, August 2015, in press.

[38] Jennifer E Mendell, Kendall D Clements, J Howard Choat, and Esther R Angert. Extreme polyploidy in a large bacterium. *Proceedings of the National Academy of Sciences*, 105(18):6730–6734, May 2008.

[39] Teemu P Miettinen, Heli K J Pessa, Matias J Caldez, Tobias Fuhrer, M Kasim Diril, Uwe Sauer, Philipp Kaldis, and Mikael Björklund. Identification of transcriptional and metabolic programs related to mammalian cell size. *Current biology : CB*, 24(6):598–608, March 2014.

[40] J M Mitchison. Growth during the cell cycle. *International review of cytology*, 226:165–258, 2003.

[41] Gautham Nair, Travis Walton, John Isaac Murray, and Arjun Raj. Gene transcription is coordinated with, but not dependent on, cell divisions during C. elegans embryonic fate specification. *Development (Cambridge, England)*, 140(16):3385–3394, August 2013.

[42] John R S Newman, Sina Ghaemmaghami, Jan Ihmels, David K Breslow, Matthew Noble, Joseph L DeRisi, and Jonathan S Weissman. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature*, 441(7095):840–846, June 2006.

[43] Zuqin Nie, Gangqing Hu, Gang Wei, Kairong Cui, Arito Yamane, Wolfgang Resch, Ruoning Wang, Douglas R Green, Lino Tessarollo, Rafael Casellas, Keji Zhao, and David Levens. c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell*, 151(1):68–79, September 2012.

[44] Sarah P Otto. The evolutionary consequences of polyploidy. *Cell*, 131(3):452–462, November 2007.

[45] J Peccoud and B Ycart. Markovian Modeling of Gene-Product Synthesis. *Theoretical Population Biology*, 48(2):222–234, October 1995.

[46] Jason H Pomerantz, Semanti Mukherjee, Adam T Palermo, and Helen M Blau. Reprogramming to a muscle fate by fusion recapitulates differentiation. *Journal of cell science*, 122(Pt 7):1045–1053, April 2009.

[47] Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi. Stochastic mRNA synthesis in mammalian cells. *PLoS Biology*, 4(10):e309, October 2006.

[48] Arjun Raj and Sanjay Tyagi. Detection of individual endogenous RNA transcripts in situ using multiple singly labeled probes. *Methods in enzymology*, 472:365–386, 2010.

[49] Arjun Raj, Patrick van den Bogaard, Scott A Rifkin, Alexander van Oudenaarden, and Sanjay Tyagi. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods*, 5(10):877–879, October 2008.

[50] Arjun Raj and Alexander van Oudenaarden. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–226, October 2008.

[51] Arjun Raj and Alexander van Oudenaarden. Single-molecule approaches to stochastic gene expression. *Annual review of biophysics*, 38(1):255–270, 2009.

[52] Simon Renny-Byfield and Jonathan F Wendel. Doubling down on genomes: polyploidy and crop plants. *American journal of botany*, 101(10):1711–1725, October 2014.

[53] K D Robertson, K Keyomarsi, F A Gonzales, M Velicescu, and P A Jones. Differential mRNA expression of the human DNA methyltransferases (DNMTs) 1, 3a and 3b during the G(0)/G(1) to S phase transition in normal and tumor cells. *Nucleic Acids Research*, 28(10):2108–2113, May 2000.

[54] Alvaro Sanchez and Ido Golding. Genetic determinants and cellular constraints in noisy gene expression. *Science*, 342(6163):1188–1193, December 2013.

[55] E E Schmidt and U Schibler. Cell size regulation, a mechanism that controls cellular RNA accumulation: consequences on regulation of the ubiquitous transcription factors Oct1 and NF-Y and the liver-enriched transcription factor DBP. *The Journal of Cell Biology*, 128(4):467–483, February 1995.

[56] Adrien Senecal, Brian Munsky, Florence Proux, Nathalie Ly, Floriane E Braye, Christophe Zimmer, Florian Mueller, and Xavier Darzacq. Transcription factors modulate c-Fos transcriptional bursts. *CellReports*, 8(1):75–83, July 2014.

[57] Alex K Shalek, Rahul Satija, Xian Adiconis, Rona S Gertner, Jellert T Gaublomme, Raktima Raychowdhury, Schraga Schwartz, Nir Yosef, Christine Malboeuf, Diana Lu, John J Trombetta, Dave Gennert, Andreas Gnirke, Alon Goren, Nir Hacohen, Joshua Z Levin, Hongkun Park, and Aviv Regev. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236–240, June 2013.

[58] Samuel O Skinner, Leonardo A Sepúlveda, Heng Xu, and Ido Golding. Measuring mRNA copy number in individual Escherichia coli cells using single-molecule fluorescent in situ hybridization. *Nature protocols*, 8(6):1100–1113, June 2013.

[59] David M Suter, Nacho Molina, David Gatfield, Kim Schneider, Ueli Schibler, and Felix Naef. Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 332(6028):472–474, April 2011.

[60] Peter S Swain, Michael B Elowitz, and Eric D Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12795–12800, October 2002.

[61] Joe Swift, Irena L Ivanovska, Amnon Buxboim, Takamasa Harada, P C Dave P Dingal, Joel Pinter, J David Pajerowski, Kyle R Spinler, Jae-Won Shin, Manorama Tewari, Florian Rehfeldt, David W Speicher, and Dennis E Discher. Nuclear lamin-A scales with tissue stiffness and enhances matrix-directed differentiation. *Science*, 341(6149):1240104–1240104, August 2013.

[62] H Tani, R Mizutani, K A Salam, K Tano, K Ijiri, A Wakamatsu, T Isogai, Y Suzuki, and N Akimitsu. Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Research*, 22(5):947–956, May 2012.

[63] Yuichi Taniguchi, Paul J Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X Sunney Xie. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–538, July 2010.

[64] Vladimir Tchernaenko, Herbert R Halvorson, Mikhail Kashlev, and Leonard C Lutter. DNA bubble formation in transcription initiation. *Biochemistry*, 47(7):1871–1884, February 2008.

[65] Tatjana Trcek, Jeffrey A Chao, Daniel R Larson, Hye Yoon Park, Daniel Zenklusen, Shailesh M Shenoy, and Robert H Singer. Single-mRNA counting using fluorescent in situ hybridization in budding yeast. *Nature protocols*, 7(2):408–419, February 2012.

[66] Jonathan J Turner, Jennifer C Ewald, and Jan M Skotheim. Cell size control in yeast. *Current biology : CB*, 22(9):R350–9, May 2012.

[67] A Tzur, R Kafri, V S LeBleu, G Lahav, and M W Kirschner. Cell Growth and Size Homeostasis in Proliferating Animal Cells. *Science*, 325(5937):167–171, July 2009.

[68] Ester Valls, Sara Sánchez-Molina, and Marian A Martínez-Balbás. Role of histone modifications in marking and activating genes through mitosis. *The Journal of biological chemistry*, 280(52):42592–42600, December 2005.

[69] Diana Y Vargas, Arjun Raj, Salvatore A E Marras, Fred Russell Kramer, and Sanjay Tyagi. Mechanism of mRNA transport in the nucleus. *Proceedings of the National Academy of Sciences of the United States of America*, 102(47):17008–17013, November 2005.

[70] Dmitri Volfson, Jennifer Marciniak, William J Blake, Natalie Ostroff, Lev S Tsimring, and Jeff Hasty. Origins of extrinsic variability in eukaryotic gene expression. *Nature*, 439(7078):861–864, December 2005.

[71] Naoharu Watanabe, Takeshi Ishihara, and Yasumi Ohshima. Mutants carrying

two sma mutations are super small in the nematode C. elegans. *Genes to Cells*, 12(5):603–609, May 2007.

[72] H Weiss-Schneeweiss, K Emadzade, T-S Jang, and G M Schneeweiss. Evolutionary consequences, constraints and potential of polyploidy in plants. *Cytogenetic and genome research*, 140(2-4):137–150, 2013.

[73] Michael L Whitfield, Gavin Sherlock, Alok J Saldanha, John I Murray, Catherine A Ball, Karen E Alexander, John C Matese, Charles M Perou, Myra M Hurt, Patrick O Brown, and David Botstein. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular biology of the cell*, 13(6):1977–2000, June 2002.

[74] Angela R Wu, Norma F Neff, Tomer Kalisky, Piero Dalerba, Barbara Treutlein, Michael E Rothenberg, Francis M Mburu, Gary L Mantalas, Sopheak Sim, Michael F Clarke, and Stephen R Quake. Quantitative assessment of single-cell RNA-sequencing methods. *Nature Methods*, 11(1):41–46, January 2014.

[75] Chia-Yung Wu, P Alexander Rolfe, David K Gifford, and Gerald R Fink. Control of transcription by cell size. *PLoS Biology*, 8(11):e1000523, 2010.

[76] J Wuarin and U Schibler. Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. *Molecular and cellular biology*, 14(11):7219–7225, November 1994.

[77] Daniel Zenklusen, Daniel R Larson, and Robert H Singer. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature structural & molecular biology*, 15(12):1263–1271, December 2008.

[78] L Zhao, C D Kroenke, J Song, D Piwnica-Worms, J J H Ackerman, and J J

Neil. Intracellular water-specific MR of microbead-adherent cells: the HeLa cell intracellular water exchange lifetime. *NMR in biomedicine*, 21(2):159–164, 2008.

[79] Jacob Zhurinsky, Klaus Leonhard, Stephen Watt, Samuel Marguerat, Jürg Bähler, and Paul Nurse. A coordinated global control over cellular transcription. *Current biology : CB*, 20(22):2010–2015, November 2010.

[80] C J Zopf, Katie Quinn, Joshua Zeidman, and Narendra Maheshri. Cell-cycle dependence of transcription dominates noise in gene expression. *PLoS Computational Biology*, 9(7):e1003161, 2013.